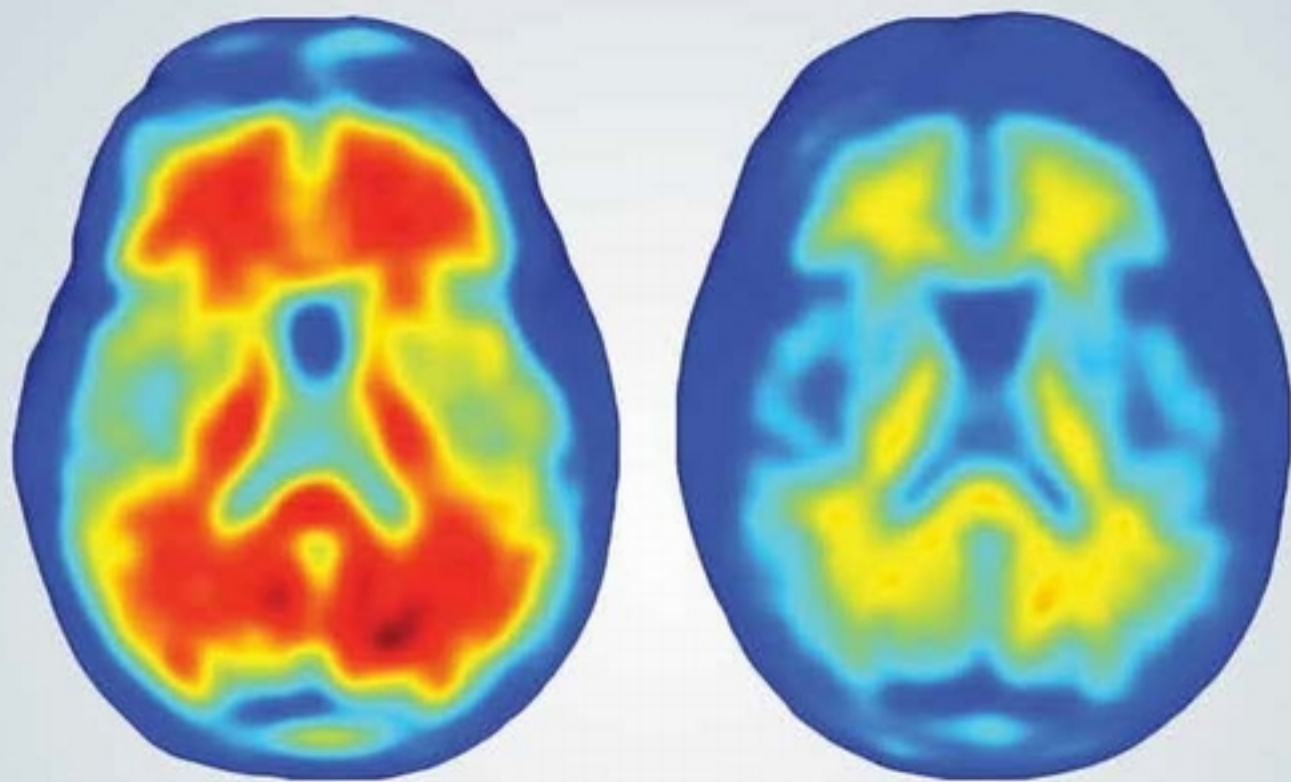


nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



TARGETING AMYLOID

Antibody aducanumab reduces Alzheimer's disease-associated amyloid in human brain **PAGES 36 & 50**

COMPUTING

DNA MEMORIES

Genomic technology
tackles big data

PAGE 22

RESEARCH MISCONDUCT

CHEATING HAPPENS

Don't ignore the fraud factor
in irreproducibility

PAGE 29

ATOMIC THEORY

SPHERES OF INFLUENCE

How John Dalton's wooden
models defined the atom

PAGE 32

NATURE.COM/NATURE

1 September 2016 £10

Vol. 537, No. 7618



THIS WEEK

EDITORIALS

WORLD VIEW The beginning of the end for digital privacy **p.9**

PHYSICS High-fives as LHC confirms pentaquark is real **p.10**



MARINE MATHS Cuttlefish count shrimps to bag bigger meals **p.11**

Zika imbalance

The US government should not redirect vital funds to work on the Zika virus at the expense of other health priorities.

Late in 2014, the US government finally put serious effort and money into combating the Ebola outbreak in West Africa. The US\$5.4-billion emergency fund approved by Congress was the largest amount of funding ever appropriated for a single international health crisis. Numerous voices — including *Nature's* — applauded the investment while warning against using the money to crush the outbreak and then fly back home. Experts warned that, without permanent improvements to Africa's health-care systems, Ebola or something worse would reappear. Health workers set out to use the funds, which were intended to last until 2019.

But the short attention span of politicians and the public has endangered these efforts. Resources intended to address the root causes of epidemics are being transferred to research on Zika, the mysterious virus that appeared in South America in 2015. Zika is not deadly and — for most people — not particularly incapacitating. But it has captured attention as the latest global health threat. Cash-strapped agencies such as the US Centers for Disease Control and Prevention have run with that theme, as have politicians eager to slam their adversaries in Congress.

In February, President Barack Obama called for Congress to authorize \$1.9 billion in emergency funds to respond to Zika, including improved surveillance, international aid for health care and vector control, and research towards a vaccine. Congress refused, so in April the White House shifted nearly \$600 million from the Ebola fund that was still being used to improve disease-response training and screening.

The July announcement that mosquitoes in Florida are transmitting Zika has redoubled calls for action. Two weeks ago, the administration directed the National Institutes of Health (NIH) to move \$34 million from its research portfolio to Zika vaccine research; \$47 million will also be transferred from other medical-service budgets. Each NIH institute — even those that do not focus on infectious disease — is

contributing about 0.14% of its budget to the Zika effort, according to numbers supplied to *Nature*.

This crosses a line. Even when one sets aside global scourges such as malaria — which affects millions of people each year and rarely draws strident calls for emergency funds — Zika is just one more virus that affects the United States. Others include West Nile virus (which has no approved human vaccine), dengue and chikungunya, as well as the seasonal and circulating influenza viruses that can kill thousands.

Taking money from much-needed research and health care to develop a vaccine against one disease itself costs lives. One analysis estimated that redirecting money to Ebola and away from common infections such as malaria and HIV caused nearly as many deaths as Ebola itself (A. S. Parpia *et al. Emerg. Infect. Dis.* <http://doi.org/bcqt>; 2016). Triaging disease research and funding is always a complex issue, and, too often, public sentiment rather than health need drives policy.

But it doesn't need to be that way. When Congress returns to session on 6 September, the US administration should insist on a permanent fund from which public-health agencies can draw, similar to that made available for natural disasters. It should also dedicate more money to international surveillance, detection and health-care systems — the sort of work that the Ebola fund is intended to support — and implement more stringent vector-control strategies to keep many viruses in check.

Each plea for emergency funds underscores just how unprepared the United States is for major health crises. The Obama administration has rightly called for permanent emergency funds and money for overall infrastructure improvement. But its willingness to sacrifice necessary research and development programmes to stick Band-Aids on the latest public-health scare erodes its credibility. When a truly deadly and pervasive pathogen appears in the United States, will there be any Band-Aids left? ■

Pachyderm plight

Analysis highlights the threat to a newly distinct species of African elephant.

Contrary to common wisdom, most researchers now accept that African elephants are actually two distinct species. On the savannah lives the huge *Loxodonta africana*, whereas the smaller, secretive *Loxodonta cyclotis* is found in the forests of central Africa.

Poaching is devastating both populations, but poaching of forest elephants should be of particular concern. Research by George Wittemyer and his colleagues indicates that most females of this species do not become pregnant for the first time until they are 23, and they

produce only 1 calf every 5 to 6 years (A. K. Turkalo *et al. J. Appl. Ecol.* <http://dx.doi.org/10.1111/1365-2664.12764>; 2016). By contrast, the savannah elephant begins breeding at 12 years of age, and typically produces young at 3- to 4-year intervals. Thus, forest-elephant populations increase in size slowly, and are at greater risk of extinction.

Wittemyer's work should spur increased focus on poaching prevention, and the study is also likely to reignite debate about the failure of the International Union for Conservation of Nature (IUCN) to recognize two different African elephant species on its extinction-risk 'red list'. The IUCN has shied away from splitting the animals into two groups, primarily over fears about what this would mean for the status of hybrids between savannah and forest animals (see go.nature.com/2bo5nx3).

But the net effect of lumping the two together is to significantly underestimate the vulnerability of the African forest elephant. At its conservation congress this week, the IUCN needs to catch up with the science and recognize the real threat of this species' extinction. ■



Preserve personal freedom in networked societies

Broad anti-discrimination laws and practices could compensate for failing data protection and technology-linked loss of privacy, says Christoph Bock.

Surveillance is no longer the prerogative of government agencies. It is privatized, decentralized — and often self-inflicted. Mobile phones trace where we go and with whom we communicate. Smartwatches measure heart rates and will soon start logging happiness and anger. The resulting data are streamed over vulnerable networks to commercial servers; they may be used by advertising companies or shared on social networks.

Current data-protection laws are not prepared for this new reality. Conceptualized in the 1970s and 80s, they were designed for a society that perceived official government databases as the main privacy risk. Their focus on centralization, parsimony and secrecy clashes with today's reality of ubiquitous personal data, deliberate sharing in social networks and all-too-frequent data leaks.

We are quick to blame naive users and careless software developers when personal data are compromised, but the truth is that prudent individual behaviour provides little protection from networked surveillance. Even if I stop using my mobile phone to navigate the digital and physical world, I will still appear in the records of the people around me.

Emerging technologies aggravate the situation. Camera drones watch us from above. Augmented-reality games such as *Pokémon Go* allow developers (or their sponsors) to control where we go in the real world. And handheld DNA sequencers will not only enable real-time monitoring of airborne pathogens (and exciting citizen-science projects), but also reveal our genetic data to anybody who can obtain our DNA.

Large data sets as substrates for computer algorithms and machine-learning technology assist our daily lives — suggesting where to eat, which book to read and how to stay healthy. But they can be used against us, for example by predicting credit risk or the likelihood of committing a crime. Such predictions can be remarkably accurate, but they struggle with unusual behaviour and often discriminate against minorities. This emergent discrimination is difficult to avoid because it is rarely hard-coded into the algorithms but arises from biased training data. People might start to 'act mainstream' just to be on the safe side — certainly not desirable for a pluralistic society.

So how can we mitigate the inherent risks that 'big data' pose for personal freedom, as billions of connected devices churn out personal data, and data protection by secrecy has become an illusion?

We must remember that data protection is a means to an end, rather than a goal in itself. We do not protect data because the data would take harm; rather, we seek to protect the rights and well-being of individuals who might be harmed by certain uses of their data. This observation could hold the key to protecting personal freedom in a world of evaporating privacy. Finding ways to tame harmful uses of personal

data would make future data leaks and unguarded data sharing less of a threat. We can distinguish between essentially financial risks, defined by damages that could be fully compensated through (potentially large) financial payments, and social risks, which affect interpersonal relationships in a way that cannot be reduced to monetary transactions.

Financial risks include higher health-insurance premiums due to genetic risk factors, or waiting longer in a service hotline because the address or a prediction algorithm indicates a low-value customer. Strong anti-discrimination and consumer-protection laws can mitigate these risks, especially when combined with protection for whistle-blowers who uncover violations, and hardship funds that provide compensation when a perpetrator cannot pay.

Social risks include shaming by friends and family over compromising video footage, or attacks over a personal opinion that has become public. Social risks are hard to tackle by legislation, as individuals are unlikely to sue family members for fair and equal treatment. Nevertheless, anti-discrimination laws help mitigate social risks by sending an authoritative message that certain types of discrimination are inappropriate, creating a spillover effect into aspects of our everyday lives not normally controlled by laws and litigation.

Strong anti-discrimination laws thus emerge as a cornerstone of personal freedom when data protection fails and secrecy is compromised by ubiquitous data sharing. The European Union's Charter of Fundamental Rights shows that such protection is legally and politi-

cally achievable, prohibiting discrimination by "sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation". The Canadian Human Rights Act also provides relatively broad protection. But the situation is much more fragmented in the United States, and insufficient in China, Japan and large parts of the developing world.

Scientists can contribute to ensuring that the loss of privacy through technology does not result in loss of personal freedom. First, they can credibly assess current and future privacy risks of new technologies and stress the need to move beyond the unsustainable concept of data protection by secrecy. Second, they should advocate for robust legal protection against discrimination around the world. Third, they should educate, advise and monitor, to make sure that facts — not fears — dominate the political debate. ■

Christoph Bock is a principal investigator at the CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences in Vienna.
e-mail: cbock@cemm.oeaw.ac.at

PRUDENT
INDIVIDUAL
BEHAVIOUR PROVIDES
LITTLE
PROTECTION
FROM NETWORKED
SURVEILLANCE.

RESEARCH HIGHLIGHTS

Selections from the
scientific literature

METABOLISM

Diet restriction makes fat brown

Very low-calorie diets — shown to boost longevity in some mammals — can turn white, energy-storing fat into beige, energy-burning fat in mice.

Calorie restriction and the accumulation of beige and brown fat have both been associated with metabolic benefits such as increased sensitivity to insulin. To look for a link between the two, Mirko Trajkovski of the University of Geneva in Switzerland and his colleagues cut the calories given to normal-weight and obese mice by 40% and found that this triggered the browning of white fat into beige fat in both types of animal.

The restricted diet also raised the levels of certain immune-system proteins called cytokines. Mice that were genetically engineered to lack responses to these cytokines did not turn fat beige in response to caloric restriction — and also did not experience many of the metabolic benefits.

Cell Metab. <http://dx.doi.org/10.1016/j.cmet.2016.07.023> (2016)

NEUROSCIENCE

Memory trick dampens phobia

Recalling fearful memories shortly before receiving psychological therapy could help people to diminish long-held fears.

Once retrieved, a memory can be disrupted before it is reconsolidated — returned to long-term storage in the brain. In a study of people with a lifelong fear of spiders, Johannes Björkstrand and

his colleagues at Uppsala University in Sweden presented volunteers with pictures of spiders to activate their fear memory. They then performed exposure therapy (repeatedly showing pictures of spiders) either 10 minutes later, during memory reconsolidation, or six hours later, after reconsolidation had finished.

In a comparison of the two groups the following day, those who were treated at 10 minutes showed reduced activation in the amygdala — a brain area that mediates fear — while

viewing pictures of spiders. They were also more likely to choose to view a picture of a spider in exchange for money. *Curr. Biol.* <http://dx.doi.org/10.1016/j.cub.2016.08.022> (2016)

PARTICLE PHYSICS

Exotic pentaquark confirmed

After multiple false detections, physicists have now confirmed in a pair of studies the existence of

subatomic particles known as 'pentaquarks'.

In the standard model, particles called baryons, which make up most of the visible matter in the Universe and include protons and neutrons, are built from three fractionally charged objects called quarks. Theorists have predicted that quarks could aggregate into larger groups and have speculated for years about the short-lived pentaquark, composed of four quarks and an antiquark. Now researchers at the LHCb experiment at



JEREMY HORNER/CORBIS/VCG/GETTY

HYDROLOGY

South Asia water supplies at risk

Groundwater supplies in northern India, Pakistan, Nepal and Bangladesh could be more endangered by contamination than by depletion.

The Indo-Gangetic Basin includes the Indus, Ganges and Brahmaputra river systems and is one of the world's most heavily used freshwater reservoirs. Previous low-resolution satellite data suggested that current exploitation rates are unsustainable. To study the region in greater detail, Alan MacDonald at the British Geological Survey in Edinburgh and his colleagues examined records from nearly 3,500 water

wells and other high-resolution data to estimate groundwater levels and quality within the top 200 metres of the aquifer. The team found that 60% of the system was plagued with high levels of salt, arsenic and other pollutants. But across 70% of the aquifer, the water table has been stable, or has even risen, from 2000 to 2012.

Groundwater quality should be monitored to provide data for policymakers, the authors suggest.

Nature Geosci. <http://dx.doi.org/10.1038/ngeo2791> (2016)

CERN's Large Hadron Collider near Geneva, Switzerland, have come up with the most convincing evidence yet for this exotic particle.

In one study, the authors reanalysed previous particle-decay data while reducing their model's assumptions. They showed at extremely high statistical significance that pentaquarks are needed to explain the data. In the second study, the researchers examined data from a particular kind of decay, finding that they are in line with predictions of decays involving pentaquarks.

Phys. Rev. Lett. <http://doi.org/bpsb>; <http://doi.org/bpsb> (2016)

ANIMAL BEHAVIOUR

Cuttlefish can count

Cuttlefish seem to be able to distinguish between large and small numbers, at least when it comes to food.

Tsang-I Yang and Chuan-Chin Chiao at National Tsing Hua University in Hsinchu, Taiwan, let pharaoh cuttlefish (*Sepia pharaonis*; pictured) in the lab choose between two chambers containing different numbers of shrimps to eat. The animals consistently selected the chamber with more shrimps, regardless of whether there was a large or small difference in prey numbers. The cuttlefish also opted for two shrimps that were smaller and easier to eat than one large shrimp. But if they were hungry, they took the bigger and trickier meal.

This shows that cuttlefish have a number sense, and

that their choice of prey is motivated by both hunger and the size of the potential reward, the authors say.

Proc. R. Soc. B 283, 20161379 (2016)

GEOCHEMISTRY

Rare mineral found on Earth

Volcanic rocks from Israel contain the first known occurrence on Earth of a titanium-rich mineral called tistarite. The discovery suggests that deep-Earth chemistry may differ from what scientists had suspected.

Until now, tistarite had been found only in a single meteorite from Mexico. A team led by William Griffin at Macquarie University in Sydney, Australia, found more of it in rocks from Mount Carmel.

Tistarite forms in chemically reducing conditions, for instance in high-hydrogen environments. The authors suggest that hydrogen or methane might percolate deep into volcanic plumbing systems, creating ultra-reducing pockets in which the unusual mineral can form.

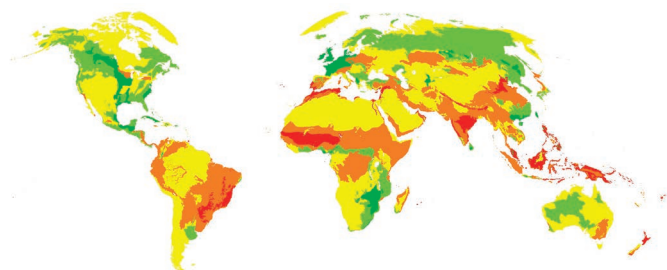
Geology <http://dx.doi.org/10.1130/G37910.1> (2016)

INFECTION

Effects of sexually spread Zika

Vaginal infection of pregnant mice by the Zika virus can cause growth restriction, brain infection and death of the fetus.

Some people have been infected by the Zika virus



Human footprint change

■ Improved ■ Slightly improved ■ Slightly degraded ■ Degraded ■ Highly degraded

through sexual activity rather than from mosquito bites. To study the effects of the virus after sexual transmission, a team led by Akiko Iwasaki at the Yale University School of Medicine in New Haven, Connecticut, developed a mouse model for vaginal transmission of the virus. They found that this mode of infection caused pathology in the fetuses in immunologically normal mothers; previous studies had suggested that Zika could not sustain long-lived infections in such animals when injected into the skin.

The results implicate the female genital tract as a particularly vulnerable site for Zika infection.

Cell <http://dx.doi.org/10.1016/j.cell.2016.08.004> (2016)

CONSERVATION

Uneven growth of human footprint

The human footprint on the global environment increased by just 9% from 1993 to 2009, even though the world's population grew by 23% and the economy by 153% during that period. However, this varied by region.

A previous study had looked at humanity's impacts on the terrestrial globe, using satellite and survey data from 1993 to quantify built environments, agricultural land, population density and other variables. To update the work, Oscar Venter at the University of Northern British Columbia in Prince George, Canada, and his colleagues compared those numbers with 2009 data.

They found that areas with the highest levels of

biodiversity, including many tropical areas, showed the fastest growth of the human footprint (pictured, in red and orange). Wealthy nations and those with strong control of corruption and high rates of urbanization showed the least growth in impacts (green).

Nature Commun. 7, 12558 (2016); *Sci. Data* 3, 160067 (2016)

NEUROSCIENCE

Protein controls brain's thermostat

A heat-sensitive protein in the brain helps to detect and regulate body temperature in mice.

Previous research had suggested that the ion channel TRPM2, which allows ions to pass across cell membranes, is involved in sensing warm temperatures. Now Jan Siemens at the University of Heidelberg in Germany and his colleagues report that the channel is expressed in a part of the hypothalamus, a brain region that helps to control body temperature. When injected with a molecule that triggers fever, mice lacking TRPM2 had higher body temperatures than control animals. The team also found that activating TRPM2-expressing neurons decreased body temperature, whereas inhibiting those neurons increased it.

The authors suggest that TRPM2 helps the brain to limit the severity of a fever.

Science <http://doi.org/bpzg> (2016)

➔ **NATURE.COM**

For the latest research published by Nature visit:

www.nature.com/latestresearch



SEVEN DAYS

The news in brief

POLICY

Union rights

A US national labour board has ruled that graduate students in the United States who work as teaching and research assistants at private universities must be recognized as employees, and therefore have a right to unionize. Graduate-student unions are already common at public institutions. The 23 August ruling relates to a case involving a group of students at Columbia University in New York City who have struggled to get their union recognized. There has been debate in recent years over the rights of graduate students, many of whom teach courses while completing their degrees.

Zika blood scans

The US Food and Drug Administration (FDA) advised US blood banks on 26 August to test all blood donations for Zika virus, in light of the virus's spread in the United States (see page 7). Thousands of US travellers have been infected with Zika virus, but since July, 29 people in south Florida have contracted it locally through mosquitoes, and the virus is expected to spread

NUMBER CRUNCH

US\$2 bn

The latest estimate of the clean-up cost of a 2014 accident at a New Mexico underground nuclear-waste repository. The sum would make the nuclear accident, in which a drum containing radioactive waste blew up, the costliest in US history.

Source: Los Angeles Times



JAMES WATT/PNNM

Obama creates largest marine park

US President Barack Obama announced the creation of the world's largest marine protected area on 26 August, with a huge expansion of the Papahānaumokuākea park in the northwest of the Hawaiian Islands. The move will take the park from its current

size of around 360,000 square kilometres to 1.5 million square kilometres. The area is home to wildlife including whales, corals, millions of seabirds and the endangered Hawaiian monk seal (*Neomonachus schauinslandi*, pictured).

to other states. Previously, the FDA recommended Zika blood screening only in states affected by the virus. Separately, Singapore has reported its first small cluster of locally transmitted cases. It joins Vietnam, Thailand and the Philippines as countries in southeast Asia that have also reported their first sporadic transmissions of the virus this year.

PEOPLE

Child-health chief

Medical geneticist Diana Bianchi will be the new head of the US National Institute of Child Health and Development (NICHD), the National Institutes of Health (NIH) announced on 25 August. She replaces Alan Guttmacher, who retired in

September 2015. As director, Bianchi will oversee the NICHD's US\$1.3-billion annual budget, which includes the Human Placenta Project and participation in a new NIH longitudinal study called Environmental Influences on Child Health Outcomes. Bianchi, who studies prenatal diagnostics, will take the helm on 31 October.

Iranian physicist

Omid Kokabee, a physicist who has been imprisoned in Iran for five years on an espionage conviction, has been granted freedom on parole, his lawyer said on 29 August. Kokabee, 34, was working on his PhD when he was jailed in Tehran in 2011. In April this year, he was moved to hospital to have kidney-cancer

surgery. He was then granted temporary medical leave and released after his friends posted bail. Kokabee has maintained his innocence and said that he was persecuted for refusing to work on a military nuclear programme in Iran. See go.nature.com/2cb5ab0 for more.

Physicist dies

US particle physicist James Cronin died on 25 August, aged 84. In 1964, with colleague Val Fitch and their collaborators, Cronin discovered anomalies in the decay of kaon particles in an accelerator experiment at the Brookhaven National Laboratory in New York. The anomalies revealed a subtle asymmetry between matter and antimatter known as CP violation. Cronin and Fitch

received a Nobel prize for their discovery in 1980. In the 1990s, Cronin became a driving force behind the Pierre Auger Observatory in Malargüe, Argentina, the largest cosmic-ray facility in the world, completed in 2004. Cronin was in the faculty of the University of Chicago in Illinois.

EVENTS

Italy earthquake

A 6.2-magnitude earthquake struck central Italy in the early hours of 24 August, killing some 290 people and devastating towns in the Apennine mountains. The quake struck 40 kilometres from L'Aquila, where a similar event killed around 300 people in 2009. The region is tectonically complex, and seismologists had expected a rupture to occur there at any time. More than 900 aftershocks occurred, impeding recovery efforts. See page 15 for more.

Airlander nosedive

The world's largest aircraft, which had a successful maiden flight in mid-August, has crash-landed on its second attempt. The 92-metre-long *Airlander 10*, which combines aeroplane and airship technology, nosedived on landing after the 100-minute test flight in Bedfordshire,



UK, on 24 August (pictured). The cockpit of the craft was damaged, but nobody was injured, said the Airlander's developer Hybrid Air Vehicles of Bedford. *Airlander 10* is intended for use in surveillance, communication, aid delivery and even passenger travel.

'No Planet B'

More than 150 Australian scientists sent an open letter on 24 August to the country's prime minister, Malcolm Turnbull, urging action on global warming. The 2015 Paris climate agreement remains unbinding, and the world's governments are "presiding over a large-scale demise of the planetary ecosystems", the scientists wrote. Citing Turnbull's 2010 statement that humanity has an obligation to the planet, the scientists called on the Australian government to do

what is required to reduce carbon emissions and coal exports. "There is no Planet B," the scientists wrote.

RESEARCH

Leprosy vaccine

India is to begin testing the world's first vaccine that exclusively targets leprosy. The disease, which is caused by the bacterium *Mycobacterium leprae*, newly affects 125,000 people in India each year — 60% of global new cases. The vaccine, developed in India, has been approved by the country's drug-regulation agency as well as the US Food and Drug Administration. According to media reports, tests will begin in a few weeks in five districts in Bihar and Gujarat, treating people who live in close contact with infected individuals. Trials have shown that infections could be reduced by 60% in 3 years.

COMING UP

4-7 SEPTEMBER

Researchers gather at the 10th Vaccine Congress in Amsterdam.

www.vaccinecongress.com

6-9 SEPTEMBER

Enthusiasts head to the British Science Festival for activities and talks.

britishsciencefestival.org

China set for Mars

The China National Space Administration is moving ahead with plans to send a rover to Mars in 2020. On 23 August, officials unveiled details of the lander, which will explore a low-latitude area in Mars's northern hemisphere. The six-wheeled probe, to be named by a public contest, is designed to operate for at least 6 months; its 13 payloads will include a ground-penetrating radar to study rock layers. Other agencies aiming to send rovers to Mars during the 2020 launch opportunity include NASA and the European Space Agency.

Robo-taxi trial

Technology company nuTonomy said on 25 August that it will start trials of self-driving taxis in Singapore, in which customers will be able to request a ride using a smartphone app. Engineers from the company, which is based in Cambridge, Massachusetts, and Singapore, will ride in the car, ready to take the wheel as needed. The joint project with the Singapore Land Transport Authority aims to launch a fully autonomous taxi service by 2018. US ride-hailing company Uber and carmaker Volvo have said that they are starting similar trials in Pittsburgh, Pennsylvania.

NATURE.COM

For daily news updates see:

www.nature.com/news

TREND WATCH

By 2085, most cities will be too hot to host the summer Olympics, according to an analysis in *The Lancet* (K. R. Smith *et al.* *Lancet* 388, 642–644; 2016). Using climate modelling and a measure of heat stress, researchers judged the suitability of cities on the basis of whether conditions would be safe to run a marathon. Looking at the Northern Hemisphere, they found 25 cities in western Europe — and just 8 elsewhere — where temperatures were likely to be less than 26°C in the shade, defined as low risk for marathon running.

CLIMATE CHANGE VERSUS THE SUMMER OLYMPICS

Most cities might be too hot to host a summer Games after 2085; western European cities may be the most suitable.



NEWS IN FOCUS

NUCLEAR POWER Plans are afoot to keep ageing plants running past 2050 **p.16**

PALAEONTOLOGY Digital scans could help unravel how famous hominin died **p.19**

NETWORKS Most mathematicians hail from just 24 scientific 'families' **p.20**

TECHNOLOGY Computing firms look to DNA to store world's data **p.22**



MASSIMO PERCOSSI/EPA



The town of Amatrice in central Italy has been devastated by the earthquake on 24 August.

SEISMOLOGY

Italian scientists shocked by earthquake devastation

In a region known to be seismically active, destruction on this scale was still a surprise.

BY ALISON ABBOTT AND
QUIRIN SCHIERMEIER

A devastating 6.2-magnitude earthquake in central Italy on 24 August that killed more than 290 people was the country's largest since a magnitude-6.3 earthquake in 2009 that hit the town of L'Aquila, about 40 kilometres away. That event killed 308 people, destroyed tens of thousands of homes and a university. Controversially, it also caused six

scientists to be put on trial for manslaughter.

Central Italy's complex geological and tectonic make-up creates a notorious quake risk. The Adria micro-plate dives beneath the Apennine mountain range from east to west, creating seismic strain. The mighty Eurasian and African plates also collide here, with the Eurasian plate moving northeast at 24 millimetres per year.

The latest quake also injured hundreds and laid waste to historic villages in the Apennine mountains, including Amatrice (see

'Epicentre of a quake'). It was a result of increased horizontal stress perpendicular to the mountain chain.

Seismologists had expected a rupture to occur near the location at any time. Still, Giulio Selvaggi, a research director at the National Institute of Geophysics and Volcanology in Rome, and one of those initially convicted of manslaughter — all six were cleared on appeal — says he was shocked by the death and destruction wreaked by last week's ►

► quake. The mountainous region around Amatrice is sparsely populated, but the final death toll may exceed that of more populated and urbanized L'Aquila.

Selvaggi seconds a public outcry over the failure of authorities to prioritize making old buildings more earthquake-resistant and notes that his team supplies earthquake maps to them. "We scientists have made a beautiful, detailed seismic hazard map, showing clearly the areas in greatest need of preventive measures," he says. "But public authorities don't take enough action."

The court case over the L'Aquila earthquake came about because a local amateur researcher claimed to have evidence of an imminent, large quake. Six scientists and one government official who had publicly dismissed the amateur's methods were accused of misinforming the public. Following an unprecedented trial, all



seven were given six-year jail sentences for manslaughter, but the scientists were cleared on appeal in 2014.

Computer scientist Paola Inverardi, who is rector of the university in L'Aquila, says the rebuilding of the university is nearly complete, and that research activities had resumed by 2012. Science in the region has also benefited from supporting initiatives following the quake, she says. One of these is the Gran Sasso Science Institute, an international graduate school founded in 2012 to inject young intellectual life into L'Aquila. It has been so successful that in June it was awarded university status.

Unlike the earthquake in L'Aquila, which was preceded by frequent, mostly low-magnitude, tremors in the surrounding area, no seismic activity was recorded before the latest earthquake. "It came out of the blue, without the preceding tremors we experienced in 'our' earthquake," says Inverardi. L'Aquila itself experienced virtually no damage, but, she says, "psychologically we were all pushed back". ■

ENERGY

Nuclear power plants prepare for old age

Efforts are afoot to keep the world's reactors running well past 2050.

BY JEFF TOLLEFSON

Sophisticated inspections are helping to pick up defects in ageing nuclear power plants before they cause trouble. In March, ultrasonic tests identified signs of wear and tear in some of the stainless-steel bolts in the reactor core of the Indian Point power plant just north of New York City. Researchers at the Electric Power Research Institute (EPRI) in Palo Alto, California, are now analysing more than a dozen of the 5-centimetre-long bolts — which secure plates that help direct water

through the radioactive core — to determine why they failed the inspection.

The analysis comes as the US Nuclear Regulatory Commission (NRC) considers whether to extend the life of Indian Point's two 40-year-old reactors for 20 more years. Opponents of the plant, including the state of New York, cite the defective bolts, a transformer fire last year and environmental and safety concerns as evidence that the facility should close.

The plant's damaged bolts are just one example of the maintenance issues facing ageing nuclear reactors around the world. The

International Atomic Energy Agency and the NRC are developing management guidelines for these facilities, but the problem may be most acute for the United States, whose fleet of 99 reactors is the oldest and largest.

The NRC has renewed the licences of 81 US reactors still in operation for another 20 years. And it presented safety guidelines in December for utilities considering renewing their licences for another 20 years. But concerns remain about the effects of time on facilities that could be in operation for 80 years (see 'Going, going, gone').

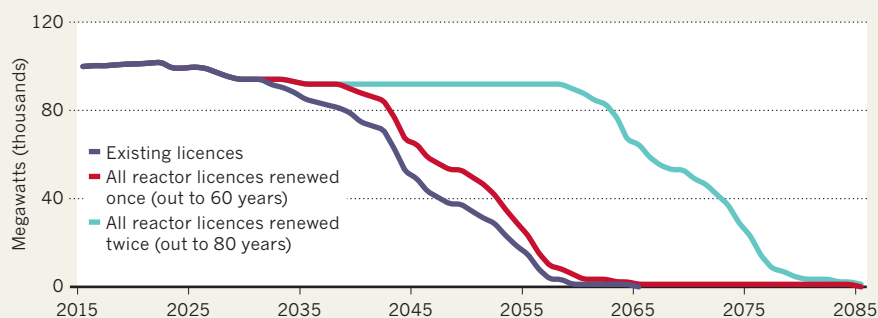
Former NRC chair Allison Macfarlane says that the industry has been struggling economically in the face of cheap natural gas, and that many nuclear power companies are investing the bare minimum when it comes to maintenance and upgrades. She would rather see a transition to newer — and safer — reactor designs than attempts to push old ones to their limits.

EXTENDING LIFETIMES

Kurt Edsinger, director of materials at the EPRI, and his team will run a battery of tests on some of the Indian Point bolts to examine fractures and assess the strength of the material. They will also analyse the effects of roughly four decades of neutron bombardment on the crystalline structure of the steel in the bolts.

GOING, GOING, GONE

Nuclear power accounts for 20% of US electricity generation, but few new reactors are being built. The following shows the total projected energy output of these plants under different licence-renewal plans.





Many US nuclear-reactor facilities are old. The Browns Ferry plant in Athens, Alabama, opened in 1974.

The study is part of a larger effort by the EPRI and the US Department of Energy to inform the industry and regulators around the world about the risks regarding ageing materials and components as nuclear power plants come up for further licence renewals.

“So far, there have been no generic show-stoppers identified that would preclude a second licence renewal,” says Kathryn McCarthy, technical director of the energy department’s Light Water Reactor Sustainability Program.

With few new reactors coming on line around the world, the longevity of existing facilities could have huge implications for the global climate. Nuclear plants currently provide 20% of the United States’ electricity — and more than half of its low-carbon power. At the global level, only hydropower provides more low-carbon power, at roughly 16% of total electricity produced, compared with nearly 11% for nuclear.

“If you maintain them and replace parts, there is no reason why nuclear plants can’t run a very long time, which is great news from a climate perspective,” says Michael Shellenberger, president of the Environmental Progress advocacy group in Berkeley, California.

Others are less sanguine. Important questions remain regarding the durability of parts that inspectors cannot see, such as underground power cables, as well as about how materials age, says Macfarlane.

Of particular concern are the concrete containment structures and steel pressure vessels at the heart of reactors, as well as the

kilometres of wires that snake through the plants. Researchers are now analysing the long-term effects of intense heat and neutron bombardment on a plant’s crucial materials down to the atomic level.

In some cases, scientists conduct accelerated-ageing experiments, in which materials are intensely irradiated to simulate 80 years of activity inside a reactor. That information can then be plugged into models that project degradation.

EARLY WARNING

The NRC’s licence-renewal process focuses on crucial infrastructure that might not be part of regular maintenance programmes. The goal is to create an inspection system that detects defects before they become a problem, says Allen Hiser, a senior technical adviser in the NRC division that handles licence renewals.

NRC officials say this is what happened at Indian Point; similar bolt defects were discovered in 1988 at a nuclear reactor in France, and the agency established inspection requirements to detect such issues in the future.

But that is not the whole story, says Dave Lochbaum, head of the nuclear-safety project at the Union of Concerned Scientists advocacy group in Cambridge, Massachusetts. The ultrasonic inspection that identified the damaged bolts at Indian Point — a technique that is now mandatory — came about only after the state of New York challenged the adequacy of visual inspections nearly a decade ago, he says.

Macfarlane remains sceptical. If the licences for current US plants are renewed for a second time, the facilities will live to be 80 years old, with nearly 100-year-old designs, she says. “We would be much better off with some of the newer reactors.” ■

“If you maintain them and replace parts, there is no reason why nuclear plants can’t run a very long time.”

ANTHROPOLOGY

Print your own 3D hominin to work out how Lucy died

Digital scans will help to test whether the famous australopithecine fell out of a tree.

BY EWEN CALLAWAY

The world's most famous fossil is now open source. 3D scans of Lucy — a 3.18-million-year-old hominin found in Ethiopia — were released on 29 August, allowing anyone to examine her arm, shoulder and knee bones and even make their own 3D-printed copies.

The scans accompany a *Nature* paper that argues that Lucy, a human relative belonging to the species *Australopithecus afarensis*, died after falling from a tree (J. Kappelman *et al.* *Nature* <http://dx.doi.org/10.1038/nature19332>; 2016). The team behind the paper also made the scans available to the public and is eager for other researchers to test the hypothesis by printing out the bones.

"It's one thing for me to describe it in detail in paper, but it's another thing to hold these things, to be able to print them out, look at them and put them together," says team leader John Kappelman, a palaeoanthropologist at the University of Texas at Austin.

His team received approval from the National Museum of Ethiopia and the country's government to make the models of Lucy public. "My sense from the Ethiopians is that Lucy is not only their national treasure, but they see her as a treasure for humankind," says Kappelman, who hopes that the country will soon release digital scans of the rest of Lucy and that other countries may follow suit with other hominin fossils.

"Coming from Ethiopia, it really is a positive step, because other countries that are hesitant may be willing to do the same thing," says Louise Leakey, a palaeontologist at Stony Brook University in New York.

But Kappelman and others say that such a move could threaten cash-strapped museums — many of them in Africa — that rely on



Lucy's arm bone undergoes a computed-tomography scan.

income generated from casts of their fossil collections to help them survive.

Lucy's digital debut was eight years in the making. Her 40%-complete remains spent 10 days in Kappelman's lab in August 2008 during a US tour. His team worked day and night to scan every one of several hundred bone fragments using a computed-tomography (CT) imager.

Close examination revealed unusual fractures: the end of her right humerus that connected to her shoulder had a series of clean breaks and compressions similar to those that orthopaedic surgeons often see in people who attempt to break a fall with an outstretched arm. Damage to Lucy's pelvis, left shoulder and knee and right ankle was also consistent with a fall from a great height. Kappelman's team

estimates that Lucy fell from a tree taller than 10 metres and died from her injuries, reaching a speed of up to 60 kilometres per hour at impact.

ARBOREAL ORIGINS

It's unclear how suited Lucy was to arboreal life. She walked upright, but she may have held onto adaptations that helped her ancestors cope in trees — although that idea is hotly debated. Kappelman's team proposes that Lucy would have slept in trees to avoid predators, yet was not as adroit there as her more-ape-like ancestors. "Here's the most famous fossil on the planet, the centre of the debate over arborealism in human evolution, and we think it's most likely she died from a fall out of tree," he says. ▶



**MORE
ONLINE**

IMAGE GALLERY



Floods, fires, Zika and a hidden portrait
go.nature.com/2boktzk

NEWS

- Academics warn of universities on the brink in South Africa go.nature.com/2bc5xam
- Meet the drones that are changing science go.nature.com/2bx9cjp
- Giant deadly ice slide baffles researchers go.nature.com/2bc63os

NATURE PODCAST



Famous hominin Lucy in shock tree fall death; antibody shows promise in treating Alzheimer's nature.com/nature/podcast

► But Marc Meyer, a palaeoanthropologist at Chaffey College in Rancho Cucamonga, California, who recently examined Lucy in Addis Ababa, is sceptical. Chimpanzees tend to break their spines when they fall from trees, says Meyer, and “Lucy’s spine does not come close to the amount of damage we would expect to see in a fatal fall”.

Lucy’s discoverers noticed her broken bones when they found her, but proposed that this had occurred after she died. Donald Johanson, the palaeoanthropologist at Arizona State University in Tempe who found Lucy in 1974, still stands by that interpretation. Broken bones such as Lucy’s are common in other nearby remains, he notes.

Kappelman is keen for others to test their theory. Digital models of portions of Lucy’s left knee and right shoulder and arm are available at eLucy.org.

But although printed bones and virtual models can be helpful, Meyer says there is no substitute for seeing a fossil in person. He found stark differences between *Ardipithecus ramidus*, a 4.4-million-year-old hominin also found in Ethiopia, and a physical cast that he studied, including several deformities not captured in the cast.

DIGITAL DOWNLOADS

Digital models of hominin fossils are rare, but a few are available. About 100 of the 1,500 remains ascribed to *Homo naledi*, uncovered in 2013 in a South African cave system, can be downloaded at MorphoSource.org, as can models of the 2-million-year-old *Australopithecus sediba* found by the same team in 2008.

AfricanFossils.org, which distributes digital models of hominin fossils for education and is headed by Leakey, contains numerous important specimens from Kenya. But the website’s models, although sufficient for 3D printing in many cases, are purposefully low in resolution, so as not to cut into income generated from making physical replicas.

Kappelman would like to see such revenue streams maintained, for instance by making lower-quality models free while charging researchers for good digital reproductions. “What has to be done is to put together a good business model that allows these museums to be able to have some sort of revenue stream off of these data,” he says.

Leakey, however, thinks that charging researchers will further limit access. She also points out that digital models can easily be pirated. “The days of keeping this content squirrelled away are gone,” she says. “Once you make a 3D model available, to control it is impossible.” ■

GENEALOGY

The ‘family trees’ of mathematics

Academic relationships hint at science, and world, history.

BY DAVIDE CASTELVECCHI

Most of the world’s mathematicians fall into just 24 scientific ‘families’, one of which dates back to the fifteenth century. The insight comes from an analysis of the Mathematics Genealogy Project (MGP), which aims to connect all mathematicians, living and dead, into family trees on the basis of teacher–pupil lineages, in particular who an individual’s doctoral adviser was¹.

The analysis also uses the MGP — the most complete such project — to trace trends in the history of science, including the emergence of the United States as a scientific power in the 1920s and the rise to dominance of different mathematical subfields.

“You can see how mathematics has evolved in time,” says Floriana Gargiulo, who studies networks dynamics at the University of Namur, Belgium, and who led the analysis.

The MGP is hosted by North Dakota State University in Fargo and co-sponsored by the American Mathematical Society. Since the early 1990s, its organizers have mined information from university departments and from individuals who make submissions regarding themselves or people they know about. As of 25 August, the MGP contained 201,618 entries. As well as doctoral advisers and pupils of mathematicians, the MGP contains details such as the university that awarded the doctorate.

Previously, researchers had used the MGP to reconstruct their own PhD-family trees, or to see how many ‘descendants’ a researcher has. Gargiulo’s team wanted to make a comprehensive analysis of the entire database and divide it into distinct families, rather than just looking at how many descendants any one person has.

After downloading the database, Gargiulo and her colleagues wrote machine-learning algorithms that cross-checked and complemented the MGP data with information from Wikipedia and from scientists’ profiles in the Scopus bibliographic database.

This revealed 84 distinct family trees with two-thirds of the world’s mathematicians concentrated in just 24 of them. The high degree of clustering arises in part because the algorithms assigned each mathematician just one academic parent: when an individual had more than one adviser, they were assigned the one

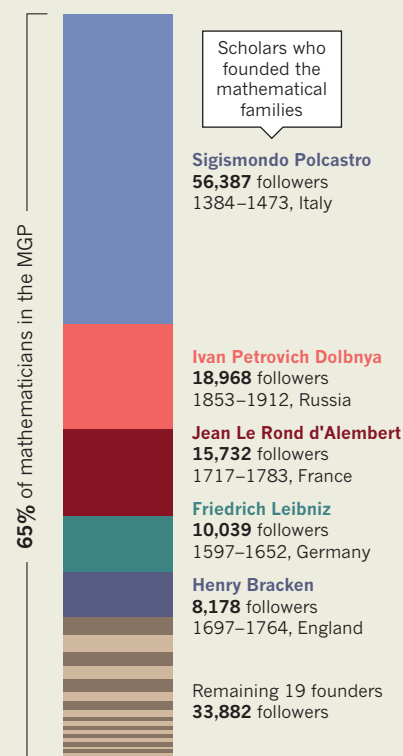
with the bigger network. But the phenomenon chimes with anecdotal reports from those who research their own mathematical ancestry, says MGP director Mitchel Keller, a mathematician at Washington and Lee University in Lexington, Virginia. “Most of them run into Euler, or Gauss or some other big name,” he says.

Although the MGP is still somewhat US-centric, the goal is for it to become as international as possible, Keller says.

Peculiarly, the progenitor of the largest family tree is not a mathematician but a physician: Sigismondo Polcastro, who taught medicine at the University of Padua in Italy in the early fifteenth century. He has 56,387 descendants according to the analysis (see ‘Mathematical clans’). The second-largest tree is one started by a Russian called Ivan Dolbnya

MATHEMATICAL CLANS

Two-thirds of mathematicians in the Mathematics Genealogy Project (MGP) belong to just 24 distinct academic families, according to an analysis that assigns ‘parenthood’ based on teacher–pupil relationships.



SOURCE: GARGIULO ET AL/MGP

► But Marc Meyer, a palaeoanthropologist at Chaffey College in Rancho Cucamonga, California, who recently examined Lucy in Addis Ababa, is sceptical. Chimpanzees tend to break their spines when they fall from trees, says Meyer, and “Lucy’s spine does not come close to the amount of damage we would expect to see in a fatal fall”.

Lucy’s discoverers noticed her broken bones when they found her, but proposed that this had occurred after she died. Donald Johanson, the palaeoanthropologist at Arizona State University in Tempe who found Lucy in 1974, still stands by that interpretation. Broken bones such as Lucy’s are common in other nearby remains, he notes.

Kappelman is keen for others to test their theory. Digital models of portions of Lucy’s left knee and right shoulder and arm are available at eLucy.org.

But although printed bones and virtual models can be helpful, Meyer says there is no substitute for seeing a fossil in person. He found stark differences between *Ardipithecus ramidus*, a 4.4-million-year-old hominin also found in Ethiopia, and a physical cast that he studied, including several deformities not captured in the cast.

DIGITAL DOWNLOADS

Digital models of hominin fossils are rare, but a few are available. About 100 of the 1,500 remains ascribed to *Homo naledi*, uncovered in 2013 in a South African cave system, can be downloaded at MorphoSource.org, as can models of the 2-million-year-old *Australopithecus sediba* found by the same team in 2008.

AfricanFossils.org, which distributes digital models of hominin fossils for education and is headed by Leakey, contains numerous important specimens from Kenya. But the website’s models, although sufficient for 3D printing in many cases, are purposefully low in resolution, so as not to cut into income generated from making physical replicas.

Kappelman would like to see such revenue streams maintained, for instance by making lower-quality models free while charging researchers for good digital reproductions. “What has to be done is to put together a good business model that allows these museums to be able to have some sort of revenue stream off of these data,” he says.

Leakey, however, thinks that charging researchers will further limit access. She also points out that digital models can easily be pirated. “The days of keeping this content squirrelled away are gone,” she says. “Once you make a 3D model available, to control it is impossible.” ■

GENEALOGY

The ‘family trees’ of mathematics

Academic relationships hint at science, and world, history.

BY DAVIDE CASTELVECCHI

Most of the world’s mathematicians fall into just 24 scientific ‘families’, one of which dates back to the fifteenth century. The insight comes from an analysis of the Mathematics Genealogy Project (MGP), which aims to connect all mathematicians, living and dead, into family trees on the basis of teacher–pupil lineages, in particular who an individual’s doctoral adviser was¹.

The analysis also uses the MGP — the most complete such project — to trace trends in the history of science, including the emergence of the United States as a scientific power in the 1920s and the rise to dominance of different mathematical subfields.

“You can see how mathematics has evolved in time,” says Floriana Gargiulo, who studies networks dynamics at the University of Namur, Belgium, and who led the analysis.

The MGP is hosted by North Dakota State University in Fargo and co-sponsored by the American Mathematical Society. Since the early 1990s, its organizers have mined information from university departments and from individuals who make submissions regarding themselves or people they know about. As of 25 August, the MGP contained 201,618 entries. As well as doctoral advisers and pupils of mathematicians, the MGP contains details such as the university that awarded the doctorate.

Previously, researchers had used the MGP to reconstruct their own PhD-family trees, or to see how many ‘descendants’ a researcher has. Gargiulo’s team wanted to make a comprehensive analysis of the entire database and divide it into distinct families, rather than just looking at how many descendants any one person has.

After downloading the database, Gargiulo and her colleagues wrote machine-learning algorithms that cross-checked and complemented the MGP data with information from Wikipedia and from scientists’ profiles in the Scopus bibliographic database.

This revealed 84 distinct family trees with two-thirds of the world’s mathematicians concentrated in just 24 of them. The high degree of clustering arises in part because the algorithms assigned each mathematician just one academic parent: when an individual had more than one adviser, they were assigned the one

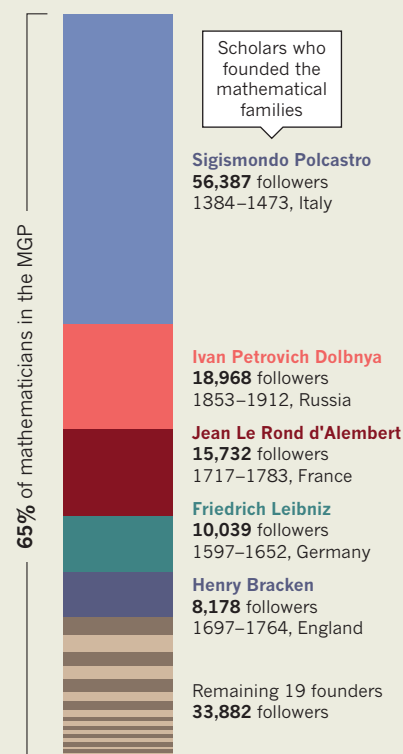
with the bigger network. But the phenomenon chimes with anecdotal reports from those who research their own mathematical ancestry, says MGP director Mitchel Keller, a mathematician at Washington and Lee University in Lexington, Virginia. “Most of them run into Euler, or Gauss or some other big name,” he says.

Although the MGP is still somewhat US-centric, the goal is for it to become as international as possible, Keller says.

Peculiarly, the progenitor of the largest family tree is not a mathematician but a physician: Sigismondo Polcastro, who taught medicine at the University of Padua in Italy in the early fifteenth century. He has 56,387 descendants according to the analysis (see ‘Mathematical clans’). The second-largest tree is one started by a Russian called Ivan Dolbnya

MATHEMATICAL CLANS

Two-thirds of mathematicians in the Mathematics Genealogy Project (MGP) belong to just 24 distinct academic families, according to an analysis that assigns ‘parenthood’ based on teacher–pupil relationships.



SOURCE: GARGIULO ET AL./MGP

in the late nineteenth century.

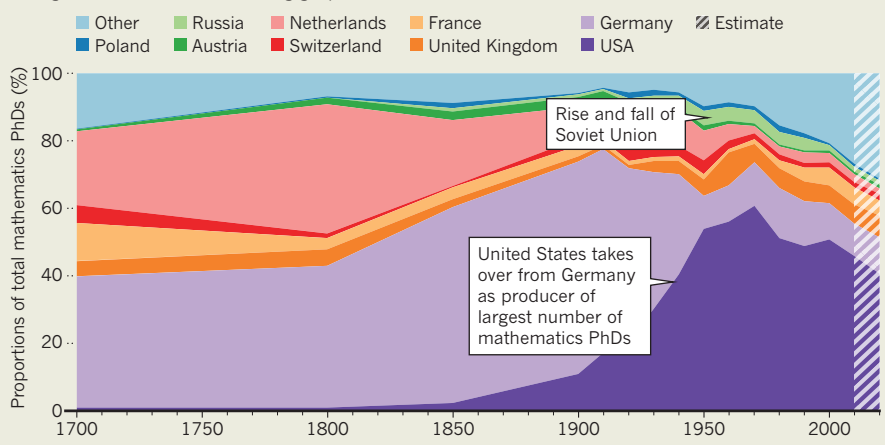
The authors also tracked mathematical activity by country, which seemed to pinpoint major historical events. Around the time of the dissolution of the Austro-Hungarian Empire, there is a decline in mathematics PhDs awarded in the region, notes Gargiulo. Between 1920 and 1940, the United States took over from Germany as the country producing the largest number of mathematics PhDs each year (see ‘Mathematics in flux’). And the ascendancy of the Soviet Union is marked by a peak of PhDs in the 1960s, followed by a relative fall after the break-up of the union in 1991.

Gargiulo’s team also looked at the dominance of mathematical subfields relative to each other. The researchers found that dominance shifted from mathematical physics to pure maths during the first half of the twentieth century, and later to statistics and other applied disciplines, such as computer science.

Idiosyncrasies in the field of mathematics could explain why it has the most comprehensive genealogy database of any discipline. “Mathematicians are a bit of a world apart,” says Roberta Sinatra, a network and data scientist at Central European University in Budapest who led a 2015 study that mapped the evolution of the subdisciplines of physics by mining data

MATHEMATICS IN FLUX

The proportion of mathematics PhDs produced by various countries changes over the centuries, tracing geopolitical trends.



from papers on the Web of Science².

Mathematicians tend to publish less than other researchers, and they establish their academic reputation not so much on how frequently they publish or on their number of citations, but on who they have collaborated with, including their mentors, she says. “I think it’s not a coincidence that they have this genealogy project.”

At least one discipline is trying to catch

up. Historian of astronomy Joseph Tenn of Sonoma State University in California plans by 2017 to launch the AstroGen project to record the PhD advisers and students of astronomers. “I started it,” he says, “because so many of my colleagues in astronomy admired and enjoyed perusing the Mathematics Genealogy Project.” ■

1. Gargiulo, F. et al. *EPJ Data Sci.* **5**, 26 (2016).
2. Sinatra, R. et al. *Nature Phys.* **11**, 791–796 (2015).

SCIENCE DIGITAL DNA

COULD THE MOLECULE KNOWN FOR STORING GENETIC INFORMATION ALSO STORE THE WORLD'S DATA?

BY ANDY EXTANCE

For Nick Goldman, the idea of encoding data in DNA started out as a joke.

It was Wednesday 16 February 2011, and Goldman was at a hotel in Hamburg, Germany, talking with some of his fellow bioinformaticists about how they could afford to store the reams of genome sequences and other data the world was throwing at them. He remembers the scientists getting so frustrated by the expense and limitations of conventional computing technology that they started kidding about sci-fi alternatives. “We thought, ‘What’s to stop us using DNA to store information?’”

Then the laughter stopped. “It was a lightbulb moment,” says Goldman, a group leader at the European Bioinformatics Institute (EBI) in Hinxton, UK. True, DNA storage would be pathetically slow compared with the microsecond timescales for reading or writing bits in a silicon memory chip. It would take hours to encode data by synthesizing DNA strings with a specific pattern of bases, and still more hours to recover that information using a sequencing machine. But with DNA, a whole human genome fits into a cell that is invisible to the naked eye. For sheer density of information storage, DNA could be orders of magnitude beyond silicon — perfect for long-term archiving.

“We sat down in the bar with napkins and biros,” says Goldman, and started scribbling ideas: “What would you have to do to make that work?” The researchers’ biggest worry was that DNA synthesis and sequencing made mistakes as often as 1 in every 100 nucleotides. This would render large-scale data storage hopelessly unreliable — unless

they could find a workable error-correction scheme. Could they encode bits into base pairs in a way that would allow them to detect and undo the mistakes? “Within the course of an evening,” says Goldman, “we knew that you could.”

He and his EBI colleague Ewan Birney took the idea back to their labs, and two years later announced that they had successfully used DNA to encode five files, including Shakespeare’s sonnets and a snippet of Martin Luther King’s ‘I have a dream’ speech¹. By then, biologist George Church and his team at Harvard University in Cambridge, Massachusetts, had unveiled an independent demonstration of DNA encoding². But at 739 kilobases (kB), the EBI files comprised the largest DNA archive ever produced — until July 2016, when researchers from Microsoft and the University of Washington claimed a leap to 200 megabytes (MB).

The latest experiment signals that interest in using DNA as a storage medium is surging far beyond genomics: the whole world is facing a data crunch. Counting everything from astronomical images and journal articles to YouTube videos, the global digital archive will hit an estimated 44 trillion gigabytes (GB) by 2020, a tenfold increase over 2013. By 2040, if everything were stored for instant access in, say, the flash memory chips used in memory sticks, the archive would consume 10–100 times the expected supply of microchip-grade silicon³.

That is one reason why permanent archives of rarely accessed data currently rely on old-fashioned magnetic tapes. This medium packs in information much more densely than silicon can, but is much slower to read. Yet even that approach is becoming unsustainable, says David Markowitz, a computational neuroscientist at the US Intelligence Advanced Research Projects Activity (IARPA) in Washington DC. It is possible to imagine a data centre holding an exabyte (one billion gigabytes) on tape drives, he says. But such a centre would require US\$1 billion over 10 years to build and maintain, as well as hundreds of megawatts of power. “Molecular data storage has the potential to reduce all of those requirements by up to three orders of magnitude,” says Markowitz. If information could be packaged as densely as it is in the genes of the bacterium *Escherichia coli*, the world’s storage needs could be met by about a kilogram of DNA (see ‘Storage limits’).

Achieving that potential won’t be easy. Before DNA can become a viable competitor to conventional storage technologies, researchers

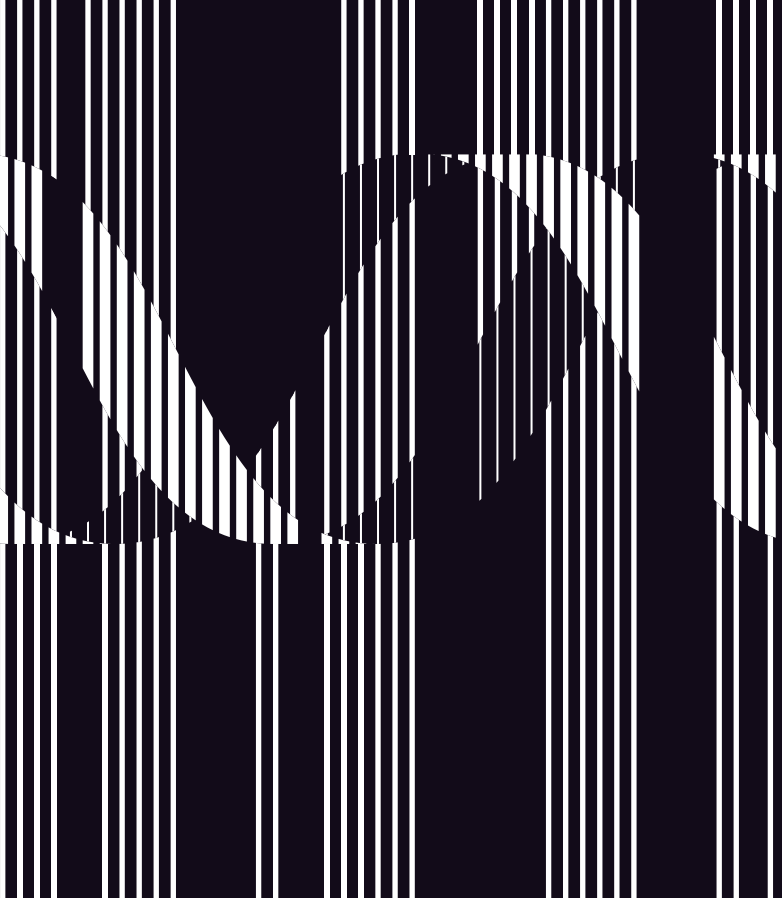


ILLUSTRATION BY WES FERNANDES/NATURE

will have to surmount a host of challenges, from reliably encoding information in DNA and retrieving only the information a user needs, to making nucleotide strings cheaply and quickly enough.

But efforts to meet those challenges are picking up. The Semiconductor Research Corporation (SRC), a foundation in Durham, North Carolina, that is supported by a consortium of chip-making firms, is backing DNA storage work. Goldman and Birney have UK government funding to experiment with next-generation approaches to DNA storage and are planning to set up a company to build on their research. And in April, IARPA and the SRC hosted a workshop for academics and industry researchers, including from companies such as IBM, to direct research in the field.

“For ten years we’ve been looking beyond silicon” for data archiving, says SRC director and chief scientist Victor Zhirnov. “It is very difficult to replace,” he says. But DNA, one of the strongest candidates yet, “looks like it may happen.”

“ONCE WE CAN GET THOSE WRITTEN ON DNA, YOU CAN STICK IT IN A CAVE AND FORGET ABOUT IT.”

LONG-TERM MEMORY

The first person to map the ones and zeroes of digital data onto the four base pairs of DNA was artist Joe Davis, in a 1988 collaboration with researchers from Harvard. The DNA sequence, which they inserted into *E. coli*, encoded just 35 bits. When organized into a 5 × 7 matrix, with ones corresponding to dark pixels and zeroes corresponding to light pixels, they formed a picture of an ancient Germanic rune representing life and the female Earth.

Today, Davis is affiliated with Church’s lab, which began to explore DNA data storage in 2011. The Harvard team hoped the application might help to reduce the high cost of synthesizing DNA, much as genomics had reduced the cost of sequencing. Church carried out the

proof-of-concept experiments in November 2011 along with Sri Kosuri, now at the University of California, Los Angeles, and genomics expert Yuan Gao at Johns Hopkins University in Baltimore, Maryland. The team used many short DNA strings to encode a 659-kB version of a book Church had co-authored. Part of each string was an address that specified how the pieces should be ordered after sequencing, with the remainder containing the data. A binary zero could be encoded by the bases adenine or cytosine, and a binary one could be represented by guanine or thymine. That flexibility helped the group to design sequences that avoided reading problems, which can occur with regions containing lots of guanine and cytosine, repeated sections, or stretches that bind to one another and make the strings fold up. They didn’t have error correction in the strict sense, instead relying on the redundancy provided by having many copies of each individual string. Consequently, after sequencing the strings, Kosuri, Church and Gao found 22 errors — far too many for reliable data storage.

At the EBI, meanwhile, Goldman, Birney and their colleagues were also using many strings of DNA to encode their 739-kb data store, which included an image, ASCII text, audio files and a PDF version of Watson and Crick’s iconic paper on DNA’s double-helix structure. To avoid repeating bases and other sources of error, the EBI-led team used a more complex scheme (see ‘Making memories’). One aspect involved encoding the data not as binary ones and zeroes, but in base three — the equivalent of zero, one and two. They then continuously rotated which DNA base represented each number, so as to avoid sequences that might cause problems during reading. By using overlapping, 100-base-long strings that progressively shifted by 25 bases, the EBI scientists also ensured that there would be four versions of each 25-base segment for error-checking and comparison against each other.

They still lost 2 of the 25-base sequences — ironically, part of the Watson and Crick file. Nevertheless, these results convinced Goldman that DNA had potential as a cheap, long-term data repository that would require little energy to store. As a measure of just how long-term, he points to the 2013 announcement of a horse genome decoded from a bone trapped in permafrost for 700,000 years⁴. “In data centres, no one trusts a hard disk after three years,” he says. “No one trusts a tape after at most ten years. Where you want a copy safe for more than that, once we can get those written on DNA, you can stick it in a cave and forget about it until you want to read it.”

A BURGEONING FIELD

That possibility has captured the imaginations of computer scientists Luis Ceze, from the University of Washington, and Karin Strauss, from Microsoft Research in Redmond, Washington, ever since they heard Goldman discuss the EBI work when they visited the United Kingdom in 2013. “DNA’s density, stability and maturity have made us excited about it,” says Strauss.

And on their return to Washington state, says Strauss, she and Ceze started investigations with their University of Washington collaborator Georg Seelig. One of their chief concerns has been another major drawback that goes well beyond DNA’s vulnerability to errors. Using standard sequencing methods, there was no way to retrieve any one piece of data without retrieving all the data: every DNA string had to be read. That would be vastly more cumbersome than conventional computer memory, which allows for random access: the ability to read just the data that a user needs.

The team outlined its solution in early April at a conference in Atlanta, Georgia. The researchers start by withdrawing tiny samples from their DNA archive. They then use the polymerase chain reaction (PCR) to pinpoint and make more copies of the strings encoding the data they want to extract⁵. The proliferation of copies makes the sequencing faster, cheaper and more accurate than previous approaches. The team has also devised an alternative error-correction scheme that the group says allows for data encoding twice as dense as the EBI’s, but just as reliable.

As a demonstration, the Microsoft–University of Washington ►

MAKING MEMORIES

DNA DATA-ENCODING SCHEMES SUCH AS THIS ONE ARE DESIGNED TO MINIMIZE ERRORS IN SYNTHESIZING AND SEQUENCING THE MOLECULE — AND THEN CORRECT ANY ERRORS THAT DO OCCUR.

TEXT TO BINARY CODE

Binary ones and zeroes represent the ASCII code for part of Shakespeare's *Sonnet 18*.

...10001000010101110011110000001001100010001...
...Thou art more lovely and more...

BINARY TO TRIPLET CODE

The binary file is mathematically converted into 'trits': the zeroes, ones and twos of a three-digit code.

...2011220200021101000202212011121010111022...

TRIPLETS TO DNA CODE

A synthesis machine creates strands of DNA using the trits as a guide. At each step, the next zero, one or two is translated to one of the three bases that differ from the base just used.

...TAGATGTGTACAGACTACGCGCAGATCGACTCGACT...

DNA FRAGMENTS

The machine makes a large number of strands with overlapping segments of 100 bases each, offset by 25, 50 or 75 bases. This guarantees four copies of each section of code, making it possible to isolate and correct errors.



STORAGE LIMITS

Estimates based on bacterial genetics suggest that digital DNA could one day rival or exceed today's storage technology.

	Hard disk	Flash memory	Bacterial DNA
Read-write speed (µs per bit)	~3,000–5,000	~100	<100
Data retention (years)	>10	>10	>100
Power usage (watts per gigabyte)	~0.04	~0.01–0.04	<10 ⁻¹⁰
Data density (bits per cm ³)	~10 ¹³	~10 ¹⁶	~10 ¹⁹

WEIGHT OF DNA NEEDED TO STORE WORLD'S DATA
~1 kg

► researchers stored 151 kB of images, some encoded using the EBI method and some using their new approach, in a single pool of strings. They extracted three — a cat, the Sydney opera house and a cartoon monkey — using the EBI-like method, getting one read error that they had to correct manually. They also read the Sydney Opera House image using their new method, without any mistakes.

ECONOMICS VERSUS CHEMISTRY

At the University of Illinois at Urbana–Champaign, computer scientist Olgica Milenkovic and her colleagues have developed a random-access approach that also enables them to rewrite the encoded data⁶. Their method stores data as long strings of DNA that have address sequences at both ends. The researchers then use these addresses to select, amplify and rewrite the strings using either PCR or the gene-editing technique CRISPR–Cas9.

The addresses have to avoid sequences that would hamper reading while also being different enough from each other to stop them being mixed up in the presence of errors. Doing this — and avoiding problems such as molecules folding up because their sequences contain stretches that recognize and bind to each other — took intense calculations. “At

the beginning, we used computer search because it was really difficult to come up with something that had all these properties,” Milenkovic says. Her team has now replaced this labour-intensive process with mathematical formulae that allow them to devise an encoding scheme much more quickly.

Other challenges for DNA data storage are scale and speed of synthesizing the molecules, says Kosuri, who admits that he has not been very bullish about the idea for that reason. During the early experiments at Harvard, he recalls, “we had 700 kB. Even a 1,000-fold increase on that is 700 MB, which is a CD”. Truly making a difference to the worldwide data archiving problem would mean storing information by the petabyte at least. “It's not impossible,” says Kosuri, “but people have to realize the scale is on the order of million-fold improvements.”

That will not be easy, agrees Markowitz. “The dominant production method is an almost 30-year-old chemical process that takes upwards of 400 seconds to add each base,” he says. If this were to remain the approach used, he adds, billions of different strings would have to be made in parallel for writing to be fast enough. The current maximum for simultaneous production is tens of thousands of strings.

A closely related factor is the cost of synthesizing DNA. It accounted for 98% of the expense of the \$12,660 EBI experiment. Sequencing accounted for only 2%, thanks to a two-millionfold cost reduction since the completion of the Human Genome Project in 2003. Despite this precedent, Kosuri isn't convinced that economics can drive the same kind of progress in DNA synthesis. “You can easily imagine markets to sequence 7 billion people, but there's no case for building 7 billion people's genomes,” he says. He concedes that some improvement in costs might result from Human Genome Project–Write (HGP–write), a project proposed in June by Church and others. If funded, the programme would aim to synthesize an entire human genome: 23 chromosome pairs containing 3.2 billion nucleotides. But even if HGP–write succeeds, says Kosuri, a human genome contains just 0.75 GB of information and would be dwarfed by the challenge of synthesizing practical data stores.

Zhirnov, however, is optimistic that the cost of synthesis can be orders of magnitude below today's levels. “There are no fundamental reasons why it's high,” he says.

In April, Microsoft Research made an early move that may help create the necessary demand, ordering 10 million strings from Twist Biosciences, a DNA synthesis start-up company in San Francisco, California. Strauss and her colleagues say they have been using the strings to push their random-access storage approach to 0.2 GB. The details remain unpublished, but the archive reportedly includes the Universal Declaration of Human Rights in more than 100 languages, the top 100 books of Project Gutenberg and a seed database. Although this is much less of a synthesis challenge than the HGP–write faces, Strauss stresses the significance of the 250-fold jump in storage capacity.

“It was time to exercise our muscle handling larger volumes of DNA to push it to a larger scale and see where the process breaks,” she says. “It actually breaks in multiple places — and we're learning a great deal out of it.”

Goldman is confident that this is just a taste of things to come. “Our estimate is that we need 100,000-fold improvements to make the technology sing, and we think that's very credible,” he says. “While past performance is no guarantee, there are new reading technologies coming onstream every year or two. Six orders of magnitude is no big deal in genomics. You just wait a bit.” ■

Andy Exantse is a freelance writer in Exeter, UK.

1. Goldman, N. *et al. Nature* **494**, 77–80 (2013).
2. Church, G. M., Gao, Y. & Kosuri, S. *Science* **337**, 1628 (2012).
3. Zhirnov, V., Zadeegan, R. M., Sandhu, G. S., Church, G. M. & Hughes, W. L. *Nature Mater.* **15**, 366–370 (2016).
4. Orlando, L. *et al. Nature* **499**, 74–78 (2013).
5. Bornholt, J. *et al. in Proc. 21st Int. Conf. Archit. Support Program. Lang. Oper. Syst.* **44**, 637–649 (ACM, 2016).
6. Hossein Tabatabaei Yazdi, S. M., Yuan, Y., Ma, J., Zhao, H. & Milenkovic, O. *Sci. Rep.* **5**, 14138 (2015).



The snakebite fight

Snakes kill tens of thousands of people each year. But experts can't agree on how best to overcome a desperate shortage of antivenom.

BY CARRIE ARNOLD

Abdulsalam Nasidi's phone rang shortly after midnight: Nigeria's health minister was on the line. Nasidi, who worked at the country's Federal Ministry of Health, learnt that he was needed urgently in the Benue valley to investigate a cluster of dying patients. People were bleeding out of their noses, their mouths, their eyes. Names of spine-chilling viruses such as Ebola, Lassa and Marburg raced through Nasidi's mind.

When he arrived in Benue, he found people splayed on the ground and tents serving

as makeshift hospital wards and morgues. But Nasidi quickly realized that the cause of the mystery illness was millions of times larger than any virus. The onset of the rainy season had brought the start of spring planting for farmers in the valley, and flooding had disturbed the resident carpet vipers (*Echis ocellatus*). Many farmers were simply too poor to buy boots — and their exposed feet became targets for the highly venomous snakes.

Nasidi wanted to help, but he found himself with limited tools. He had only a small

amount of antivenom with which to neutralize the toxin — and it quickly ran out. Once the hospital exhausted its supply, people stopped coming. No one knows how many people were killed. In an average year, hundreds of Nigerians die from snakebite, and that rainy season, which started in 2012, was far from average.

Snakebites are a growing public-health crisis. According to the World Health Organization, around 5 million people worldwide are bitten by snakes each year; more than 100,000 of them die and as many as 400,000 endure amputations

MATTIAS KLUM/NGS

Bites from venomous snakes such as the Jameson's mamba (*Dendroaspis jamesoni*) are a public-health crisis.

and permanent disfigurement. Some estimates point to a higher toll: one systematic survey concluded that in India alone, more than 45,000 people died in 2005 from snakebite¹ — around one-quarter the number that died from HIV/AIDS (see 'The toll of snakebite'). "It's the most neglected of the world's neglected tropical diseases," says David Williams, a toxinologist and herpetologist at the University of Melbourne, Australia, and chief executive of the non-profit organization Global Snakebite Initiative in Herston.

Many of those bites are treatable with existing antivenoms, but there are not enough to go around. This long-standing problem became international news in September 2015, when Médecins Sans Frontières (MSF, also known as Doctors Without Borders) announced that the last remaining vials of the antivenom Fav-Afrique, used to treat bites from several of Africa's deadliest snakes, were about to expire. The French pharma giant Sanofi Pasteur in Lyons had decided to cease production in 2014. MSF estimates that this will cause an extra 10,000 deaths in Africa each year — an "Ebola-scale disaster", according to Julien Potet, a policy adviser for MSF in Paris. Yet, because most of those affected by snakebites are in the poorest regions of the world, the issue has been largely ignored.

SPOTLIGHT ON SNAKES

In May, however, the crisis was discussed for the first time at the annual World Health Assembly meeting in Geneva, Switzerland. The world's handful of snakebite specialists gathered in a small conference room in the Palais des Nations — although they shared concern over the problem, they were split about how to solve it. Many want to use synthetic biology and other high-tech tools to develop a new generation of broad-spectrum antivenoms. Others argue that existing antivenoms are safe, effective and low cost, and that the focus should be on improving their production, price and use. "From the physician perspective, patient care and public health comes before anything new," says Leslie Boyer, who directs an institute dedicated to antivenom study at the University of Arizona, Tucson.

The debate mirrors those around many other developing-world challenges, from improving agriculture to providing clean drinking water. Do people need high-tech solutions, or can cheaper, lower-tech remedies do the job? The answer is simple to Jean-Philippe Chippaux, a physician working on snakebite for the French Institute of Research for Development in Cotonou, Benin. "We have the ability to fix this problem now. We just lack the will to do it," he says.

Every December, Williams sees snakebite victims flood into the Port Moresby General Hospital in Papua New Guinea. Nearly all of

them were bitten by the taipan (*Oxyuranus scutellatus*), one of the world's deadliest snakes, which emerges at the start of the rainy season. The venom stops a victim's blood from clotting, paralyzes muscles and leads to a slow, agonizing death. It seems a far cry from Australia, where Williams is based. "There's this incredible suffering just 90 minutes away from the modern world," he says.

Yet Williams knows that these people are the lucky ones. The hospital ward, which might be treating as many as eight taipan victims at any time, is often the only place in the country with antivenom drugs. Without them, some 10–15% of all snakebite victims die; with them, just 0.5% do. The situation is reflected around the world. "Many countries don't want to admit that they have such a primeval-sounding problem," Chippaux says.

The method used to make antivenom has changed little since French physician Albert Calmette developed it in the 1890s. Researchers inject minuscule amounts of venom, milked from snakes, into animals such as horses or sheep to stimulate the production of antibodies that bind to the toxins and neutralize them. They gradually increase doses of venom until the animal is pumping out huge amounts of neutralizing antibodies, which are purified from the blood and administered to snakebite victims.

Across much of Latin America, government-funded labs typically produce antivenoms and distribute them free of charge. But in other areas, especially sub-Saharan Africa, these life-saving medications are too often out of

reach. Many governments lack the infrastructure or political will to purchase and distribute antivenom. Bribery and corruption often jack up the price of an otherwise inexpensive drug from a typical wholesale cost of US\$18 to \$200 per vial to a retail cost between \$40 and \$24,000 for a complete treatment, according to a 2012 analysis². Not all hospitals and clinics can afford the antivenom, and some won't risk buying it because their patients either can't pay for it or won't, because they doubt that it really works.

"THERE'S THIS INCREDIBLE SUFFERING JUST 90 MINUTES AWAY FROM THE MODERN WORLD."

With no reliable market for the medicines, some pharmaceutical companies have halted production. Sanofi Pasteur stopped making Fav-Afrique because, at an average retail price of around \$120 per vial, it just couldn't sell enough to make production worthwhile. A total of 35 government or commercial manufacturers produce antivenom for distribution around the world, but only 5 now make the drugs for sub-Saharan Africa. In the absence of medicines, snakebite victims have been known

to drink petrol, electrocute themselves or apply a poultice of cow dung and water to the bite, says Tim Reed, executive director of Health Action International in Amsterdam. But there are also problems with the drugs themselves, says Robert Harrison, head of the Alistair Reid Venom Research Unit at the Liverpool School of Tropical Medicine, UK. They often have a limited shelf life and require continuous refrigeration, which is a problem in remote areas without electricity. And many are effective against just one species of snake, so clinics need an array of medicines constantly on hand. (A few, such as Fav-Afrique, combine antibodies to create a broad-spectrum product.)

Venoms from spiders and scorpions typically have only one or two toxic proteins; snake venoms can have more than ten times that amount. They are a "pandemonium of molecules", says Alejandro Alagón, a toxinologist at the National Autonomous University of Mexico in Mexico City. Researchers do not always know which proteins in this toxic soup are the damaging ones — which is why some think that smarter biology could help.

OLD PROBLEM, NEW SOLUTION

Ten years ago, teams led by Harrison and José María Gutiérrez, a toxinologist at the University of Costa Rica in San José, began parallel efforts to create a universal antivenom for sub-Saharan Africa using 'venomics' and 'antivenomics'. The aim is to identify destructive proteins in venoms using an array of techniques, ranging from genome sequencing to

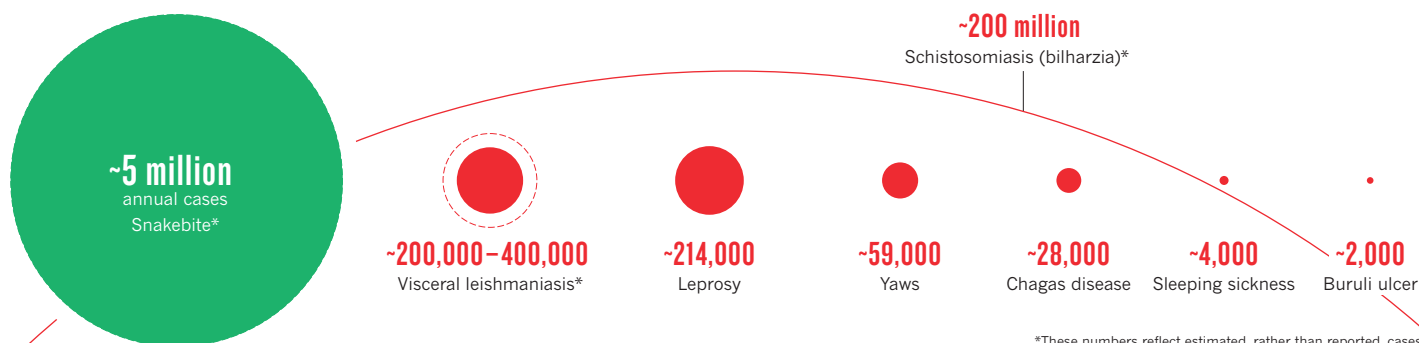
mass spectrometry, and then find the specific parts, known as epitopes, that provoke an immunological response and are neutralized by the antibodies in antivenom drugs. The ultimate goal is to use the epitopes to produce antibodies synthetically, using cells rather than animals, and develop antivenoms that are effective against a wide range of snake species in one part of the world.

The scientists have made slow but steady progress. Last year, Gutiérrez and his colleagues separated and identified the most toxic proteins from a family of venomous snakes known as elapids (Elapidae). By combining information about the abundance of each protein and how lethal it is to mice, the team created a toxicity score to indicate how important it was to neutralize a protein with antivenom, a first step towards making the treatment³.

In March this year, a Brazilian team reported that they had gone further, designing short pieces of DNA that encode key toxic epitopes

THE TOLL OF SNAKEBITE

Snakebite affects more people each year than many other neglected tropical diseases, and often causes death, disability or disfigurement. The issue receives little attention: data are scarce and the condition mostly strikes the world's poorest regions.



SOURCE: WORLD HEALTH ORGANIZATION

in the venom of the coral snake (*Micrurus corallinus*), a member of the elapid family⁴. Mice were injected with the DNA using a technique that enabled some to generate antibodies against coral-snake venom, and the group enhanced the mice's immune responses by injecting them with synthetic antibodies manufactured in bacterial cells. These and other advances led Harrison to estimate that the first trials of new antivenoms in humans could be just three or four years away. But with so few researchers working on the problem, a paucity of funding and the biological complexity of snake venoms, he and others admit that this is an optimistic prediction.

Despite the growing literature on antivenoms, Alagón and Chippaux aren't convinced that the approach will help. Alagón estimates that newly developed antivenoms would need to be priced at tens of thousands of dollars per dose to be financially viable to produce, and that no biotech or pharma company would manufacture one without substantial government subsidies. Compare that, he says, to the rock-bottom price of many existing antivenoms. "You can't get cheaper than that," he says. "We can make an entire lot of antivenoms in one day using technology that's been available for 80 years."

Finding someone to produce new medications might be a greater challenge than actually developing them, Williams acknowledges: governments or non-governmental organizations (NGOs) will almost certainly have to step in to help to defray the development costs. But he argues that now is the time to research alternative approaches. These could "revolutionize the treatment of snakebite envenoming in the next 10–15 years", Williams says.

THE ROOM WHERE IT HAPPENED

All these tensions, brewing for nearly a decade, came to a head at the Geneva meeting in May. Around 75 scientists, public-health experts and health-assembly delegates crowded around three long tables in a third-floor conference room at the United Nations Headquarters. Spring rain pelted the tall windows.

Lights were dimmed, and then the screams of a toddler filled the room. A short documentary

co-produced by the Global Snakebite Initiative told the story of a girl bitten by a cobra whose parents carried her for days over rocky roads in Africa to find antivenom. They arrived in time — the girl survived — but she lost the use of her arm. Her sister had already died after a bite from the same snake.

Convincing attendees of the scale of the problem was the meeting's primary goal; how to solve it came next. For 90 minutes, scientists and NGOs made short, impassioned speeches laying out the scope of the issue and the variety of problems that they faced. At the centre of each presentation was the same message: we need more antivenom.

But the meeting was strained. Chippaux and representatives of the African Society of Venomology were disappointed and angry that so few Africans had been invited to speak, even though the continent is where antivenom shortages are most acute. "Our voice, our issues, were completely overlooked," Chippaux says. Seated at the front of the room, group members whispered and gestured frantically to each other, and Chippaux barely managed to keep them from storming out.

They argue that the current antivenom shortage stems from Africa's reliance on foreign companies and governments for its drugs, and that the only solution lies in building up infrastructure in Africa to produce its own high-quality antivenom. Alagón views antivenomics as a dangerous diversion. "It's distracting many brilliant minds and resources from improving antivenoms using existing technology," he says. "Perhaps by 2050 this will be the standard technique, but the problem is now."

Williams and Gutiérrez take a middle ground. They feel that the problem requires attacks on all fronts. As well as innovation, Gutiérrez calls for existing manufacturers to step up the production of current drugs.

There are signs of this happening already. Latin America has a long history of producing antivenoms both for its own needs and for those of countries around the world, and even before Sanofi Pasteur announced that it would cease production of Fav-Afrique, Costa Rica, Brazil and Mexico were testing antivenoms

for different parts of Africa. One product, EchiTAB-Plus-ICB, is produced by Costa Rica and effective against a range of African viper species; it completed clinical trials in 2014 and is now available for use. Several other antivenoms are expected to be ready in the next two years. The drugs should be affordable: government labs in Costa Rica have already indicated that they will not seek to make money from the antivenoms, just recoup their expenditures.

But beyond that, the way forward remains murky. Williams knows that the World Health Assembly meeting was just a start. Inevitably, more meetings will be needed to produce a concrete action plan. But the discussion still gave him and some others a renewed sense of hope that the international community is beginning to take snakebite seriously — momentum they hope to build on by banging away at the topic at conferences and in the media.

Boyer says that whatever solution the snakebite field decides on, the most important thing is to "break the cycle of antivenom failure in Africa". Doing that requires building trust from governments, health-care workers and the public that the drugs are safe and effective, that clinics will have antivenom on hand, and that people will be able to afford treatment. "Without that, you've got nothing," Boyer says. Educating local clinics on how to care for snakebite victims and administer treatments in a timely manner would also go a long way towards preventing deaths.

Speaking of the devastation he saw in Benue, Nasidi says that something as simple as providing boots for poor farmers would have helped to prevent much of the suffering and death that he witnessed. It's perhaps the ultimate in low-tech methods in snakebite protection: shielding vulnerable human skin. ■

Carrie Arnold is a writer based near Richmond, Virginia.

1. Mohapatra, B. et al. *PLoS Negl. Trop. Dis.* **5**, e1018 (2011).
2. Brown, N. I. *PLoS Negl. Trop. Dis.* **6**, e1670 (2012).
3. Laustsen, A. H., Lohse, B., Lomonte, B., Engmark, M. & Gutiérrez, J. M. *Toxicon* **104**, 43–45 (2015).
4. Ramos, H. R. et al. *PLoS Negl. Trop. Dis.* **10**, e0004484 (2016).

COMMENT

CHEMISTRY What John Dalton did for science **p.32**

REDUCTIONISM Abstract art and experimental science compared **p.33**

REPRODUCIBILITY A call to hallmark labs that produce replicable work **p.34**

PEER REVIEW We've let the perfect become the enemy of the good **p.34**



ILLUSTRATION BY DAVID PARKINS



Stop ignoring misconduct

Efforts to reduce irreproducibility in research must also tackle the temptation to cheat, argue
Donald S. Kornfeld and Sandra L. Titus.

The history of science shows that irreproducibility is not a product of our times. Some 350 years ago, the chemist Robert Boyle penned essays on “the unsuccessfulness of experiments”. He warned readers to be sceptical of reported

work. “You will meet with several Observations and Experiments, which... may upon further tryal disappoint your expectation.” He attributed the problem to a ‘lack of skill in the scientist and the lack of purity of the ingredients’, and what would today be

referred to as inadequate statistical power.

By 1830, polymath Charles Babbage was writing in more cynical terms. In *Reflections on the Decline of Science in England*, he complains of “several species of impositions that have been practised in science”, namely “hoaxing, forging, trimming and cooking”.

In other words, irreproducibility is the product of two factors: faulty research practices and fraud. Yet, in our view, current initiatives to improve science dismiss the second factor. For example, leaders at the US National Institutes of Health (NIH) stated in 2014: “With rare exceptions, we have no evidence to suggest that irreproducibility is caused by scientific misconduct”¹. In 2015, a symposium of several UK science-funding agencies convened to address reproducibility, and decided to exclude discussion of deliberate fraud.

To dismiss the role of research misconduct is mistaken and unfortunate. At best, ignoring deliberate misconduct in efforts to reduce irreproducibility is a wasted opportunity, like tilling a field without clearing it of rocks. At worst, it permits destructive behaviour to persist and flourish.

SCALE OF EVIDENCE

Only 10–12 individuals are found guilty by the US Office of Research Integrity (ORI) each year. That number, which the NIH used to dismiss the role of research misconduct¹, is misleadingly low, as numerous studies show. For instance, a review² of 2,047 life-science papers retracted from 1973 to 2012 found that around 43% were attributed to fraud or suspected fraud. A compilation of anonymous surveys³ suggests that 2% of scientists and trainees admit that they have fabricated, falsified or modified data. And a 1996 study⁴ of more than 1,000 post-docs found that more than one-quarter would select or omit data to improve their chances of receiving grant funding.

Admittedly, many causes of irreproducibility do not involve dishonesty. The NIH has promoted responsible research for 25 years by funding studies on research integrity, creating educational resources and backing the ORI.

Nonetheless, we contend that when scientific leaders minimize “hoaxing, forging, trimming and cooking” as contributors to irreproducibility, they choose to ignore the problem rather than confront it. This ►

► mechanism is what psychiatrists term denial, when an individual faces what they believe to be an insoluble problem. Deliberate misconduct is a reality that government funders can and must address. In 2012, an article in this journal declared that “the time is right to confront misconduct”. We agree; it is even more urgent now. We recommend five key approaches (see ‘Preventing misconduct’).

TARGETED REMEDIES

In the 1990s, the NIH mandated that all of the trainees it funds must receive a course on the responsible conduct of research. Not surprisingly, it failed in its goal of reducing research misconduct⁵ — which it defines as fabrication, falsification or plagiarism. Presumably, the ethics proscribing such practices are established long before people enter science. Instead, we propose interventions to address the psychological factors that motivate individuals to commit misconduct, depending on their role in the research hierarchy.

Those found guilty of misconduct by the ORI fall into three categories in roughly equal measure: trainees, support staff and senior scientists. Each has its own motivations⁶.

Trainees. Many trainee missteps can be traced to a fear of failure and a lack of quality mentorship. One study⁷ of trainees who were found guilty of misconduct revealed that 62% of their mentors had not established adequate procedures, such as providing clear rules on data ownership and recording, safety, materials transfer or scheduling regular meetings, and 73% had not reviewed trainees’ raw data. A survey⁸ at a major US cancer centre found that nearly one-third of 140 trainees felt pressure to “prove” a mentor’s hypothesis, even though results did not support it.

Some trainees who commit misconduct are perfectionists and are unable to cope with failure. Mentors should intervene with perspective, encouragement and even referral to counselling. They should assure trainees that there are respected careers outside a tenure-track appointment. Instead, junior scientists report that they are treated as cheap labour; their professional development is a low priority.

Funders should craft policies to ensure that mentors act as an adviser, teacher and role model, and should limit the number of trainees per mentor by discipline. Each year, trainees should be required to complete anonymous questionnaires evaluating their mentors, and results should be sent to funding agencies as well as to research deans.

Institutions should reward mentors for outstanding performance and provide adequate training. When justified, mentors should be held responsible for misconduct by their trainees and appropriately sanctioned.

Support staff. Workers such as laboratory

REMEDIES

Preventing misconduct

To diminish the threat that misconduct poses to science, scientists and society:

- Authorities should acknowledge that deliberate misconduct is an important contributor to irreproducibility.
- Mentors should be evaluated to assure quality; those who contribute to misconduct should be penalized.
- Institutions and government agencies should have procedures to protect whistle-blowers from retaliation.
- Senior faculty members who are found guilty of misconduct should face severe penalties.
- Institutions that fail to establish and follow policies and processes to prevent misconduct should be sanctioned.

technicians, phlebotomists and data collectors represent about one-third of the individuals annually found guilty by the ORI of deliberate misconduct. They may have falsified data to boost their income or reduce their workload in response to an investigator’s unrealistic productivity goals.

Treating support staff as valued members of a team could go a long way. They should be made aware of the study’s goals, and how invalid publications harm scientific progress and patient care.

Senior researchers. Established scientists would be less likely to commit misconduct if they were more concerned about being detected and punished. Currently, they conclude that the risk is low: few cases are referred to the ORI and few of their colleagues want to be enmeshed in a conflict.

More than 80% of faculty members say that they would be reluctant to report potential misconduct for fear of being ostracized and damaging their own reputations⁹. One ORI study found that 47 out of 68 people who reported misconduct experienced an adverse consequence. Clearly, concerns about making allegations are justified.

Well-articulated policies are key to helping whistle-blowers come forward. So too is a well-trained research integrity officer (RIO), ideally a respected faculty member or administrator. The potential whistle-blower must be confident that the institution’s RIO and its policies will protect them from retaliation.

Institutions. Research centres should build a culture and infrastructure that encourages integrity. For example, peers can emphasize their commitment to robust data in everyday interactions and by supporting random audits; and data systems can date-stamp and

track who accesses files to protect them from manipulation. Leaders should make it clear that they will not tolerate misconduct and that perpetrators will suffer severe consequences.

An effective RIO is crucial for leading the educational and enforcement effort and in building trust in the integrity of the institution. Unfortunately, studies have shown that many RIOs are poorly trained and do not manage allegations and investigations of research misconduct effectively. Perhaps most significantly, they might fail to adequately prepare and protect the whistle-blower¹⁰. The RIO must be selected thoughtfully and provided with sufficient authority and support.

Any institution that receives US federal research funds should be required to have at least one designated, trained and certified RIO who has been assessed by the ORI. Moreover, research funds should not be released to an institution that cannot demonstrate current certification.

Institutions that fail to establish and execute policies to assure integrity should be held responsible when misconduct occurs. For example, in July 2014, Iowa State University agreed to repay US\$496,000 and forego \$1.4 million in grants after one of its researchers was found guilty of fraud. However, this penalty, as well as a prison sentence for the fraudster, happened only because a senator intervened. That should not be necessary.

Government officials should be prepared to pursue repayments. The threat of such penalties should have a chilling effect on investigators contemplating research misconduct, and motivate institutions to establish and implement policies that reflect their commitment to institutional integrity.

We believe that these system-wide interventions are essential to have an impact on the irreproducibility produced by research misconduct. ■

Donald S. Kornfeld is professor emeritus of psychiatry and special lecturer at the College of Physicians and Surgeons, Columbia University, New York City, USA. **Sandra L. Titus** is former director of intramural research at the US Office of Research Integrity, Maryland, USA.
e-mails: dsk3@cumc.columbia.edu; sandra.titus@yahoo.com

1. Collins, F. S. & Tabak, L. A. *Nature* **505**, 612–613 (2014).
2. Fang, F. C., Steen, R. G. & Casadevall, A. *Proc. Natl Acad. Sci. USA* **109**, 17028–17033 (2012).
3. Fanelli, D. *PLoS ONE* **4**, e5738 (2009).
4. Eastwood, S., Derish, P., Leash, E. & Ordway, S. *Sci. Eng. Ethics* **2**, 89–114 (1996).
5. Antes, A. L. et al. *Acad. Med.* **85**, 519–526 (2010).
6. Kornfeld, D. S. *Acad. Med.* **87**, 877–882 (2012).
7. Wright, D. E., Titus, S. L. & Cornelison, J. B. *Sci. Eng. Ethics* **14**, 323–336 (2008).
8. Mobley, A., Linder, S. K., Braeuer, R., Ellis, L. M. & Zwilling, L. *PLoS ONE* **8**, e63221 (2013).
9. Titus, S. L. *Account. Res.* **21**, 9–25 (2013).
10. Bonito, A. J. et al. *Account. Res.* **19**, 308–328 (2012).

IN RETROSPECT

A New System of Chemical Philosophy

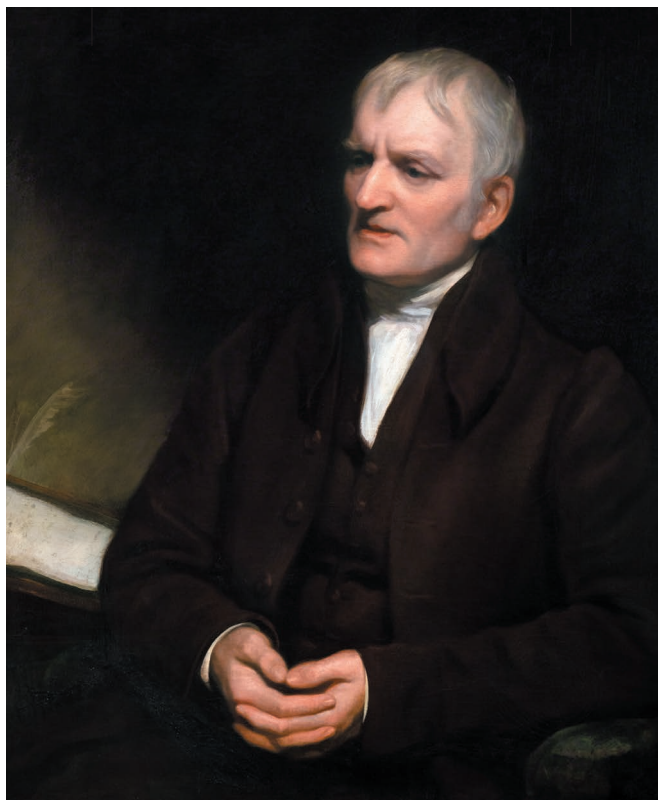
Philip Ball reflects on the work of John Dalton, father of modern atomic theory.

Visual metaphors are often essential in science when you can't see what you're studying. The English chemist John Dalton, born 250 years ago, illustrated his atomic theory using wooden spheres (pictured), drilled with holes for pins that enabled them to be linked into clusters. But there are hazards to such mental props. By the 1880s, students were so familiar with the spheres that one (taught by prominent advocate of atomic theory Henry Enfield Roscoe) declared: "Atoms are round bits of wood invented by Mr Dalton."

Today, the atoms Dalton proposed in his seminal *New System of Chemical Philosophy* (1808) are routinely revealed by microscopy and crystallography. They are corralled in electromagnetic traps, pushed around like marbles using scanning probe microscopes, even manufactured and monitored one at a time in superheavy forms using particle accelerators. No one mistakes them for bits of wood.

Neither did Dalton. He articulated the ancient idea that matter is built from fundamental particles in a way that aligned it with the quantitative principles of chemical reaction elucidated in the late eighteenth century. Those macroscopic rules, he said, stemmed from the systematic combination of microscopic bodies: solid, massy and hard, as Isaac Newton had put it in a phrase Dalton was fond of quoting.

Yet in a sense, even by the 1880s, atoms were still not much more than Dalton's model spheres. Because they remained unobserved, several leading scientists refused to accept their reality, among them physicist Ernst Mach and chemist Wilhelm Ostwald. Some considered atoms no more than an heuristic convenience: a crutch that the mind could use to make sense of chemical transformations. That is why, despite Roscoe's misgivings that Dalton's wooden balls might mislead students, the balls had a valuable role. They showed how visualizing an entity can help to cement the concept even while direct



John Dalton, painted in 1835 by Thomas Philips.

evidence is elusive. It is a risky strategy to assert the physical reality of something not yet observed (will dark matter really be particulate?). But without such an image, a theory can seem little more than metaphysics.

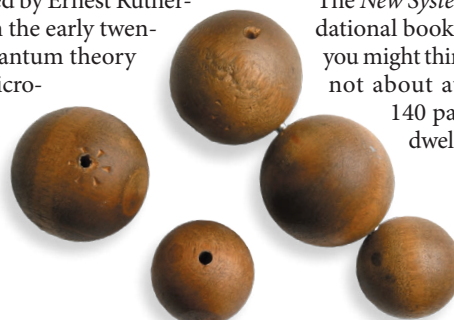
It is traditional to locate Dalton's *New System of Chemical Philosophy* as a step — perhaps the greatest — in a long road to modern atomic theory that began with the ancient Greek atomists Leucippus and Democritus in the fifth century BC, and ended with the nuclear atoms proposed by Ernest Rutherford and Niels Bohr in the early twentieth century, then quantum theory and scanning probe microscopes. The "philosophy" in Dalton's title signified something closer to a scientific theory than to the abstract reasoning

it tends to connote today. Yet his book also represents an important juncture for the philosophy of science. It spoke to whether science should be based on empiricism or explanatory hypothesis — a question that had exercised Newton and Robert Boyle in the seventeenth century. There was nothing new in Dalton's idea of atomistic matter; the question was whether to treat this as a useful conjecture or as a reality. Antoine Lavoisier, whose work on the proportions of chemical combination was crucial to Dalton, had no time for such questions. Lavoisier insisted that meditating on "ultimate particles" was metaphysical — and fruitless.

So how did Dalton, a modest teacher educated in Cumbrian village schools and excluded from Oxford and Cambridge for his Quakerism, take an imaginative leap that eluded distinguished professors? Even if we admit some of the fairy dust of "genius" into an explanation, we shouldn't discount Dalton's wide reading — from Boyle and Newton to Claude Louis Berthollet and Humphry Davy. He

also paid careful attention to the quantitative details of experiments by the likes of his friend, Mancunian chemist William Henry, and Lavoisier. Dalton presented his atomistic theory to the Manchester Literary and Philosophical Society, of which he was secretary, between 1803 and 1805. Some of his papers were published in the society's memoirs, but he was urged to present them as a book, as he put it, in "the interests of science, and his own reputation."

The *New System* is one of those foundational books that doesn't say what you might think it should. It is mostly not about atoms at all. The first 140 pages or so of Volume 1 dwell on heat and its effects,



The spheres that Dalton used to demonstrate atomic theory.

IAN DIGNALL/ALAMY

MANCHESTER MUSEUM OF SCIENCE & INDUSTRY/SSPL

A New System of Chemical Philosophy

JOHN DALTON

R. Bickerstaff: 1808.

whereas Volume 2 is a detailed account of inorganic chemical compounds. Dalton's atomic theory is confined to the five-page final chapter of the first volume. Here, he explains that the fixed stoichiometries of chemical reactions — so much of element A combines with so much of B — can be rationalized by supposing that the constituent atoms unite into “compound atoms” of simple ratios, such as 1:1 or 1:2. The point is most famously and eloquently made in a plate that shows sketches of these unions. An “atom” of water comprises one atom each of hydrogen and oxygen; an atom of ammonia is a 1:1 union of hydrogen and nitrogen (Dalton uses Lavoisier's term, “azote”, for nitrogen).

The proportions are wrong — chemist Jöns Jakob Berzelius corrected many in the following two decades. And in 1813, he proposed an alphabetical representation (for example, H^2O [sic]) in place of Dalton's pictorial balls. Dalton, with the conservatism common to trailblazers, declared this “horrificing”, saying that the symbols “cloud the beauty and simplicity of the atomic theory”. His displeasure might have contributed to a stroke in 1837.

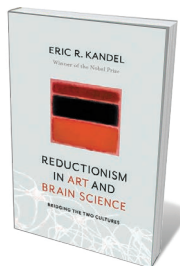
The *New System* is not a new theory of chemistry. Among other things, it offers no explanation for why atoms react. Roscoe put his finger on it when he said that the significance of Dalton's theory was his proposal that each type of atom has a unique mass. That made sense of the quantities in which elements were found to combine, and offered the first general and fundamental distinction between one element and the next — what eventually became embodied in the idea of atomic number.

Yet it is the idea of atoms as the indivisible units of matter that stuck in the mind, because readers could see them on the page. Dalton didn't intend his pedagogical diagrams of atomic unions — “compound atoms”, or molecules as we'd now say — to be taken too literally. There's no inkling in his book of molecular shape; the arrangements of atoms in binary, ternary and other unions are purely notional, and when Dalton draws “water particles” packed into the crystalline forms of ice, they too are spheres.

All the same, visual representation of atoms was surely the precondition for the emergence of a concept of molecular structure, with atoms in fixed spatial relationships, in the mid-nineteenth century. Something of this kind would surely have appeared whether or not Dalton had “invented” atoms as wooden balls — but that innovation was more eloquent than its inventor anticipated. ■

Philip Ball is a writer based in London. His latest book is *The Water Kingdom*. e-mail: p.ball@btinternet.com

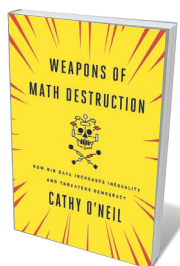
Books in brief



Reductionism in Art and Brain Science: Bridging the Two Cultures

Eric R. Kandel COLUMBIA UNIVERSITY PRESS (2016)

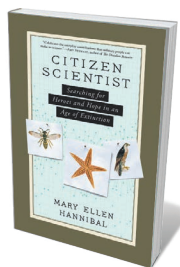
The sea-slug studies of Nobel-prizewinning neuroscientist Eric Kandel — which reveal the link between memory and synaptic connection — are models of reductionist science. In this intriguing treatise, Kandel finds methodological similarities in abstract art. By reducing image to colour, form or line, artists such as Piet Mondrian stimulated the brain's “top-down processing” in the viewer, encouraging ‘active seeing’. Kandel deconstructs this intricate dance between perceiver and perceived by way of recent neuroscience findings and deft analyses of seminal artworks.



Weapons of Math Destruction

Cathy O'Neil CROWN (2016)

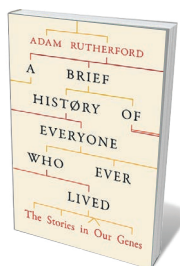
While working as a Wall Street analyst during the 2008 crash, data scientist Cathy O'Neil realized how maths can fuel social problems. Her propulsive study reveals many models that are currently “micromanaging” the US economy as opaque and riddled with bias. These algorithmic overlords can taint policing and court sentences with racial profiling, and exacerbate unemployment rates in poor communities. In an era when many people uncritically applaud the power of big data, O'Neil argues for the dark side of the deluge to be tackled through algorithm audits, transparency and legal reform.



Citizen Scientist: Searching for Heroes and Hope in an Age of Extinction

Mary Ellen Hannibal THE EXPERIMENT (2016)

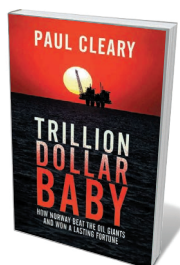
In this inside story on citizen science and biodiversity loss, Mary Ellen Hannibal meshes interviews with front-line scientists such as James Estes (*Nature* **533**, 318–319; 2016) with her own stints monitoring California wildlife. Inspired by the likes of marine biologist Ed Ricketts (*Nature* **516**, 326–328; 2014), she records starfish die-offs, meets the geeks who track deforestation, and plans a web-based supercommunity of citizen scientists to counter what many are calling the sixth great extinction. A cogent call to action.



A Brief History of Everyone Who Ever Lived: The Stories in Our Genes

Adam Rutherford WEIDENFELD & NICOLSON (2016)

Fifteen years ago, the first sequence and analysis of the human genome was published (E. S. Lander *et al.* *Nature* **409**, 860–921; 2001). A monumental surge in genetics followed. Science writer and broadcaster Adam Rutherford rides that tide and traces its effects, first focusing on how genetics has enriched and in some cases upset our understanding of human evolution, then examining the revelations of recent findings, such as deep flaws in the concept of race. Although digressive in the chapters on deep history, Rutherford unveils the science with elegance.



Trillion Dollar Baby

Paul Cleary BITEBACK (2016)

Norway's government pension fund could hit US\$1 trillion in just four years. In this crisp economic history stretching back more than four decades, journalist Paul Cleary charts how this middle-income Scandinavian country ensured that 90% of the cash flow from vast oil discoveries accrued to its government. But despite its record of pragmatic fair-mindedness, Norway's eagerness to excavate environmentally sensitive reaches of the Arctic shows how its forward planning fails when it comes to climate change. **Barbara Kiser**

Correspondence

Fallacy of perfection harms peer review

Voltaire wrote in 1772, “the best is the enemy of the good”, warning against the fallacy that something is worthless if it is not perfect — a sentiment that seems common in scientific peer review today.

The history of science has taught us that most progress has come from exploring flawed hypotheses and imperfect models. We must always strive for the better study, the better model, the better analysis. As experienced reviewers, however, we contend that seeking ultimate perfection is not the same as accepting nothing less here and now. Scientific progress depends on such compromise — provided that potential caveats are recognized.

If a model is the most technically and ethically feasible approach available, and is better than random guessing, then it has some merit in advancing knowledge. Useful developments in biology, for example, have come from *in vitro* systems that do not reflect *in vivo* conditions, and from animal models that do not necessarily predict human disorders.

The aim should be to utilize models, despite their imperfections, while continuing to improve them. It is unrealistic to hold progress in science to standards of perfection and certainty: progress is usually incremental and iterative.

James C. Zimring
BloodworksNW; and University of Washington, Seattle, USA.

Steven L. Spitalnik
Columbia University, New York, USA.
jzimring@bloodworksnw.org

Hallmark labs with a replicability record

I am concerned that the tension between good research practice and scientific success is rising, despite recent efforts to shore up replicability (see *Nature* <http://doi.org/bpmf>; 2016).

As others have noted,

high-quality research should start in the lab, by validating cell lines and reagents, for example, and end with serious, meticulous review. That rarely happens because it is time-consuming, and time is every scientist's worst enemy — particularly for young researchers who face stiff competition in the scientific job market.

To resolve this conflict, we need to work out how to change the incentive system so that it fosters a culture of good, responsible research. Reproducibility could be underpinned by a strict set of rules — including, say, systematic use of power analysis and sample-size estimation. To promote compliance and to counter any negative effect on productivity, and hence on competition for funding, labs with a record of high-quality research could be accredited with an international certificate of approval. An independent, not-for-profit organization might be responsible for awarding such certificates.

Mattia Andreoletti
European Institute of Oncology, Milan, Italy.
mattia.andreoletti@ieo.eu

Stuck between a rock and a hard place

We question the basis of July's ruling by an international tribunal that the disputed Spratly Islands (Nan-sha Islands) in the South China Sea are merely rocks (see *Nature* 535, 334–335; 2016).

According to the United Nations Convention on the Law of the Sea, which is signed by 178 countries, an island is defined by three elements: it is a natural formation; it lies in the same territory and economic zone as the surrounding sea; and it can sustain human habitation or have an economic life of its own (see go.nature.com/2bj4sit).

One of the disputed islands, Taiping Island, fulfils all three criteria, having its own freshwater resources (see go.nature.com/2biujnv; in Chinese). There

is also human habitation and economic activity on several of the other islands.

In our view, the international court seems to have interpreted an island's third defining element as requiring no external resources to sustain human settlement. If that were the case, the Maldives, Singapore and Hong Kong should probably be considered as rocks — they bring in gas and must obtain much of their fresh water by import and desalination or from rain water.

Yingchao Hu, Liangjun Hu
Northeast Normal University, Changchun, China.
hulj068@nenu.edu.cn

Rate oceans' capital to help achieve SDGs

Goal 14 of the United Nations' Sustainable Development Goals (SDGs) is dedicated to conserving and using the oceans and their resources for sustainable development. We suggest that a 'gross marine product' (GMP) index — a measure of the oceans' natural capital — would be invaluable for achieving this goal.

The seas provide us with food, materials, livelihoods and recreation. Managing these ecosystem services effectively can help us to eradicate poverty, develop sustainable economies and adapt to global environmental changes. Yet international-resource experts and national strategies still focus largely on goods and services delivered by terrestrial ecosystems (see go.nature.com/2bcqjr0).

A GMP index would provide a measure of marine ecosystem goods and services on a national or global scale, derived from estimates for individual oceans. More international research will be necessary to underpin these estimates. The results would inform decision-makers, the private sector and the public on how they could help to achieve

goal 14, as well as the 60 targets across most of the 17 SDGs that are relevant to the sustainable development of coastal zones. An integrated programme that measures, monitors and assesses the health of human-ocean systems should oversee their sustainability.

Yonglong Lu*
Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing, China.
yllu@rcees.ac.cn

*On behalf of 5 correspondents (see go.nature.com/2biddcz for full list).

Could Pokémon Go boost birding?

In a week when the game *Pokémon Go* topped 15 million downloads, I had a salutary reminder that urban humans risk losing touch with nature — with possible negative implications for the future of fieldwork in conservation and ecology (see also *Nature* 535, 323–324; 2016).

As I set out to go birdwatching in Queensland's rainforest, my 14-year-old daughter grabbed her smartphone to search for rare Pokémon in every nearby park, beach and town. The Pokémon are an extremely speciose group that undergo continuous evolution and have particular ecological needs. Embedded in nature by an augmented reality, they hold the same naturalistic delight for my daughter as a cassowary (*Casuarus casuarinus*) does for me.

At the end of my day, I had counted three 'lifers' (my first sightings of *Platalea regia*, *Entomyzon cyanotis* and *Nectarinia jugularis*) and my daughter had spotted 30 Pokémon. I was delighted when she asked me about a bird that appeared beside a Pidgey on her screen. It was a real laughing kookaburra (*Dacelo novaeguineae*).

Fabio de Oliveira Roque
Federal University of Mato Grosso do Sul, Brazil.
roque.eco@gmail.com

Verifying quantum superpositions at metre scales

ARISING FROM T. Kovachy *et al.* *Nature* **528**, 530–533 (2015); doi:10.1038/nature16155

Although the existence of quantum superpositions of massive particles over microscopic separations has been established since the founding of quantum mechanics, the maintenance of superposition states over macroscopic separations is a subject of modern experimental tests. Kovachy *et al.*¹ report on applying optical pulses to place a freely falling Bose–Einstein condensate into a superposition of two trajectories that separate by an impressive distance of 54 cm before being redirected towards one another. When the trajectories overlap, a final optical pulse produces interference with high contrast, but random phase, between the two wave packets. Contrary to ref. 1, we argue that the observed interference is consistent with, but does not prove, the claim that the spatially separated atomic ensembles were in a quantum superposition state; therefore, the persistence of such superposition states remains experimentally unestablished. There is a Reply to this Brief Communication Arising by Kovachy, T. *et al.* *Nature* **537**, <http://dx.doi.org/10.1038/nature19109> (2016).

The authors of ref. 1 equate the observation of interference with the existence of a phase-coherent quantum superposition between the separated atomic samples. However, Anderson's hypothetical experiment², which involves two independently produced, 'non-communicating' volumes of superfluid helium, emphasizes a distinction between interference and phase coherence. He pointed out that connecting the two volumes by a narrow orifice would result in a Josephson current and that the relative phase determined from the Josephson relation would have a random value. It is impossible, even in principle, for any measurement on a single realization of the set-up to determine whether this phase was established before or after the Josephson current was produced².

Such thought experiments have been realized in the laboratory, demonstrating interference between two independently generated light beams³ and between two independently produced Bose–Einstein condensates⁴. In both these examples, the spatially separated, indistinguishable quantum objects—photons in one case, sodium atoms in the other—had no defined quantum coherence between them. Each sample could have interacted with its own local environment and experienced uncorrelated perturbations therefrom. Yet, in each repetition of the experiment high-contrast interference was observed, whereas the phase of the interference was irreproducible between repetitions. The same behaviour is observed by Kovachy *et al.*

Phase-coherent quantum superposition states are characterized by first-order coherence. First-order coherence measures the expectation value of a product of two field operators, $\langle \psi_A^\dagger \psi_D \rangle$. In a many-body system, this expectation value appears in the off-diagonal element of the one-body reduced density matrix. Such coherence is measured by a two-slit interference experiment: the quantum fields emanating from two points, A and D, are allowed to interfere. The presence of first-order coherence, that is, of quantum superposition states, is indicated by an interference pattern with a determinate phase. Although it is possible that the random-phase interference observed by Kovachy *et al.* is caused by only technical imperfections in their optical pulses, their observation is also consistent with the lack of first-order coherence and of coherent quantum superpositions.

By contrast, second-order coherence measures the expectation value of a product of four field operators, $\langle \psi_A^\dagger \psi_B^\dagger \psi_C \psi_D \rangle$, and is an element within the two-body density matrix. Second-order coherence is

indicated by the fact that the interference pattern produced by two quantum fields, for example, those emanating from points A and D, is the same as that between two other quantum fields, for example, those emanating from points B and C. In the experiment of Kovachy *et al.*, the points A and B correspond to locations within the gas on the upper interferometer path and the points C and D to positions within the gas on the lower path. Their observation that the interference phase in one portion of the gas is equal to that in another portion of the gas demonstrates the existence of second-order, but not first-order, coherence.

Therefore, we assert that the experiment of Kovachy *et al.* does not demonstrate the existence of quantum superposition states of massive particles over metre length scales. The second-order coherence observed in the experiment is immune to perturbations that are common across the sub-millimetre length scale of each of the spatially separated clouds, but that differ arbitrarily over the metre-scale distance between the two paths of their atomic interferometer. Such perturbations can arise from technical imperfections or intrinsic atomic interactions. In addition, and directly relevant to the claims made by Kovachy *et al.*, these perturbations would arise from exotic effects proposed in theories of continuous spontaneous localization or gravitationally induced decoherence^{5,6}. Indeed, if these exotic localization effects localize particles with metre-scale resolution and also respect the indistinguishability of identical quantum particles, then they could collapse the states onto an incoherent mixture that has a definite mass in each arm and thus determine the exact number of atoms in each interferometer path. Such localization would completely eliminate first-order coherence between the two interferometer paths, so that the one-body density matrix becomes that of a mixed state⁶. The state produced by such localization is a 'quantum superposition' only insofar as it is composed of identical bosons, the wavefunctions of which must be symmetric under particle exchange. Yet, as in Anderson's thought experiment², the two atomic wave packets will still show high-contrast interference, with an interference phase that is random between experiments^{7,8}. We note that our argument contradicts the claim by Nimmrichter and Hornberger⁶ that single-shot measurements of atom interferometers serve to test their phenomenological model for the decay of macroscopic quantum superposition states.

The second-order coherence observed by Kovachy *et al.* does demonstrate that each of the separated Bose–Einstein condensates remains coherent over its sub-millimetre size during the 1-s time of propagation. Matter-wave coherence over similar timescales and length scales has been observed previously, as summarized in extended data figure 3 and extended data table 1 of ref. 1. In those previous works, the existence of a determinate phase is confirmed by comparing the phases of two well-separated atomic interferometers, allowing for the elimination of common-mode technical noise. The observations of Kovachy *et al.* do rule out exotic effects that would cause the position of each and every individual atom to be independently measured with metre-scale resolution. However, such effects would violate the principle of the indistinguishability of identical particles, and are thus implausible (as discussed in ref. 6).

Verification of a quantum mechanical superposition requires the measurement of a determinate phase to distinguish a pure quantum state from a statistical mixture of several pure states. Once information

about the phase is lost, whether owing to measurement noise, interactions with the environment, or a fundamental source of decoherence, no further measurement can distinguish between a quantum superposition and a mixed state. Without determinate phase information, the system in ref. 1 is consistent with being in a statistical mixture of interferometer states.

If the system examined by Kovachy *et al.* does indeed retain quantum coherence over long timescales and length scales, then evidence for such coherence could be obtained either by better phase stabilization of the optical pulses or, if that is impractical, by operating two well-separated interferometers that share the same optical pulses⁹. These improvements would enable the impressive technical advance in atom interferometry reported in ref. 1 to become a test of quantum physics at long length scales.

D. M. Stamper-Kurn^{1,2}, G. E. Marti³ & H. Müller¹

¹Department of Physics, University of California, Berkeley, California 94720, USA.

²Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA.

email: dmsk@berkeley.edu

³JILA, National Institute of Standards and Technology and University of Colorado, Boulder, Colorado 80309, USA.

Received 11 March; accepted 29 June 2016.

1. Kovachy, T. *et al.* Quantum superposition at the half-metre scale. *Nature* **528**, 530–533 (2015).
2. Anderson, P. W. in *The Lesson of Quantum Theory* (eds de Boer, J. *et al.*) 23–34 (Elsevier, 1986).
3. Magyar, G. & Mandel, L. Interference fringes produced by superposition of two independent maser light beams. *Nature* **198**, 255–256 (1963).
4. Andrews, M. R. *et al.* Observation of interference between two Bose condensates. *Science* **275**, 637–641 (1997).
5. Bassi, A. *et al.* Models of wave-function collapse, underlying theories, and experimental tests. *Rev. Mod. Phys.* **85**, 471–527 (2013).
6. Nimmrichter, S. & Hornberger, K. Macroscopicity of mechanical quantum superposition states. *Phys. Rev. Lett.* **110**, 160403 (2013).
7. Javanainen, J. & Yoo, S. M. Quantum phase of a Bose–Einstein condensate with an arbitrary number of atoms. *Phys. Rev. Lett.* **76**, 161–164 (1996).
8. Naraschewski, M. *et al.* Interference of Bose condensates. *Phys. Rev. A* **54**, 2185–2196 (1996).
9. Snadden, M. J. *et al.* Measurement of the Earth's gravity gradient with an atom interferometer-based gravity gradiometer. *Phys. Rev. Lett.* **81**, 971–974 (1998).

Author Contributions All authors contributed to the study and to the writing of this Brief Communication Arising.

Competing Financial Interests Declared none.

doi:10.1038/nature19108

Kovachy *et al.* reply

REPLYING TO D. M. Stamper-Kurn, G. E. Marti & H. Müller *Nature* **537**, <http://dx.doi.org/10.1038/nature19108> (2016)

In the accompanying Comment¹, Stamper-Kurn *et al.* assert that our observation of interference contrast in a half-metre-scale atom interferometer² does not prove the existence of macroscopic quantum superpositions and, hence, does not test quantum mechanics at long length scales. Moreover, they imply that intrinsic atomic interactions or technical imperfections could prevent the application of our work to future differential measurements. In response, we argue: (i) that in standard quantum mechanics there is no known mechanism in our system that prohibits its use in future differential measurement applications; (ii) that our experiment tests quantum mechanics in that it constrains any modifications that would reduce contrast in an interferometer with arms that propagate over widely separated trajectories; and (iii) that, using a standard definition of superposition, our observation of interference results from quantum superposition at the half-metre scale. In particular, we argue that quantum superposition is a more general concept than first-order coherence.

We operated our atom source with a condensate fraction of approximately 50%. The atom source has a coherence length of only 2×10^{-6} m, substantially smaller than the spatial extent of the atom cloud. This short coherence length arises from imperfections in the magnetic lensing, the lattice launch, and the interactions of the atoms with the Bragg laser beams. Coherence between the two interferometer arms is established by the initial beam-splitter pulse, at which time the ratio of the interaction matrix element (U) to the Bragg-transition Rabi frequency (J) is $U/J \approx 10^{-8}$, which rules out interaction-based effects during the beam-splitter pulse³. The atomic density is no larger than about 10^{10} cm^{-3} during the interferometer sequence, which is dilute enough to prevent dephasing due to mean-field shifts (the mean-field shift is approximately 0.1 Hz). Under these conditions, standard quantum mechanics rules out evolution into the state described in ref. 1. Furthermore, we know of no technical noise sources that would lead

to the emergence of such states; all known technical noise sources, such as residual spontaneous emission, are associated with momentum exchange that modifies the structure of the atomic states and reduces contrast. Therefore, there is no known mechanism that would prohibit the utilization of the acceleration sensitivity inferred from the large arm separation in differential measurement applications, such as using dual species interferometry for a test of the equivalence principle⁴.

When evaluating the degree to which our experiment constrains a particular hypothetical modification of quantum mechanics, it is important to consider disturbances to the states of individual atoms—for example, due to momentum exchange that fundamentally alters the structure of the many-atom state as it propagates. In the case of momentum exchange, the large spatial separation directly translates into an increased sensitivity to this spurious heating. A spurious momentum kick $\hbar q$ (where \hbar is the reduced Planck constant and q is the wave number associated with the momentum kick) that occurs midway through the interferometer is associated with a wave-packet phase shift of $[m(v + \hbar q/m)^2/2 - mv^2/2]T/\hbar \approx qL$, where m is the atomic mass, v is the velocity separation, T is the drift time and $L \approx vT$ is the wave-packet separation. Even momentum kicks as small as $q \approx 2\pi/L$ (corresponding to wavelength of about L) result in phase shifts of around 2π , which, if they occur inhomogeneously, result in reduced contrast. Modifications that add only overall phase noise are not ruled out by our results.

We would like to clarify our use of the word ‘superposition’. In ref. 2, following Feynman and others, we adopted the nomenclature that interference—whether or not there is a determinate phase—necessarily results from superposition (see, for example, refs 5 and 6). This view of superposition is illustrated by the Pfleegor–Mandel experiment⁷, which tracks the build-up of an interference pattern from two independent laser beams one photon at a time. A standard interpretation of these

experiments is that interference results from superposition between the sources, which is revealed during the detection process⁷.

**T. Kovachy¹, P. Asenbaum¹, C. Overstreet¹, C. A. Donnelly¹,
S. M. Dickerson¹, A. Sugarbaker¹, J. M. Hogan¹ & M. A. Kasevich¹**

¹Department of Physics, Stanford University, Stanford, California
94305, USA.

email: kasevich@stanford.edu

1. Stamper-Kurn, D. M., Marti, G. E. & Müller, H. Verifying quantum superpositions at metre scales. *Nature* **537**, <http://dx.doi.org/10.1038/nature19108> (2016).

2. Kovachy, T. *et al.* Quantum superposition at the half-metre scale. *Nature* **528**, 530–533 (2015).
3. Javanainen, J. & Ivanov, M. Y. Splitting a trap containing a Bose-Einstein condensate: atom number fluctuations. *Phys. Rev. A* **60**, 2351–2359 (1999).
4. Hogan, J. M., Johnson, D. M. S. & Kasevich, M. A. Light-pulse atom interferometry. In *Proc. Int. School Phys. Enrico Fermi* Vol. 168 (eds Arimondo, E. *et al.*) 411–447 (IOS Press, 2009).
5. Feynman, R. P., Leighton, R. B. & Sands, M. *The Feynman Lectures on Physics* Vol. III, Ch. 1–4 (Pearson, 1965).
6. Cohen-Tannoudji, C., Diu, B. & Laloë, F. *Quantum Mechanics* Vol. 1, Ch. III.E (Hermann and John Wiley & Sons, 1977).
7. Pfleegor, R. L. & Mandel, L. Interference of independent photon beams. *Phys. Rev.* **159**, 1084–1088 (1967).

doi:10.1038/nature19109

Questioning Holocene community shifts

ARISING FROM S. K. Lyons *et al.* *Nature* **529**, 80–83 (2016); doi:10.1038/nature16447

Demonstrating changes in the structure of plant and animal communities over geological time scales and linking these changes to human impacts in the Holocene epoch would be an important contribution to the fields of ecology and conservation biology¹. Lyons *et al.*² claim to provide such evidence based on a decrease in the proportion of spatially aggregated species pairs, using co-occurrence data from assemblages spanning the past 300 million years. However, we suggest that apparent flaws in their predictions, assumptions, methods and interpretations undermine this claim, and we question the conclusion that the structure of communities has fundamentally changed during the Holocene. There is a Reply to this Brief Communication Arising by Lyons, S. K. *et al.* *Nature* **537**, <http://dx.doi.org/10.1038/nature19111> (2016).

Lyons *et al.*² calculated the proportion of aggregated species pairs over the total number of significantly aggregated and segregated species pairs for all assemblages. Their conclusions are based on a linear segmented regression of the proportion of aggregated pairs against time. We believe that a linear model cannot account for the estimate errors following a binomial distribution. The Gaussian model makes the assumption that the variance of proportions is constant. Considering the residuals of the linear Gaussian model as a function of the number of species pairs, we observe that the variance decreases when the number of species pairs increases (S1). Indeed, in this study², a proportion of 50% of aggregated species could have been calculated either based on one aggregated and one segregated species pair, or based on 100 aggregated and 100 segregated pairs. The reliability of the estimate is clearly not the same. In total, 44% of the proportions are based on 5 or less species pairs from assemblages with several thousand random species pairs. Accounting for the number of species pairs using a generalized linear model (GLM; binomial error distribution, logit link), we found that the proportion of aggregated species pairs decreased indeed over time (GLM, $\chi^2_{(1,99)} = 229$, $P < 0.0001$), yet the segmentation did not reveal any breakpoint at $-6,000$ years (Fig. 1a).

Commenting on shifts in community structure in a meaningful way involves comparing the comparable—that is, communities of the same taxonomic group within the same broad geographical area over time. However, Lyons *et al.*² estimated the proportion of aggregated species pairs in 101 assemblages over a time span of 300 million years with numerous confounding factors such as taxonomic group, number of species, temporal extent or spatial grain. For example, community structure in 290-million-year-old plant assemblages with data spanning 16 million years, in which some species in a ‘community’ are unlikely to have actually co-existed, was compared with 10-year-old mammal assemblages. This is not inherently wrong, but the confounding variables should be included in the model. Yet, Lyons *et al.*² did not add these predictors nor did they test for potential interactions among them.

Instead, they used univariate correlations to test for an effect of four continuous variables on the proportion of aggregated species pairs, omitting 39 ‘modern’ assemblages (less than 100 years old) from these tests. Without these modern data, their linear model would no longer show any breakpoint, and in addition there would be no significant decline in the proportion of aggregated species pairs over time ($F_{(1,60)} = 1.72$, $P = 0.19$, analysis of variance (ANOVA)).

Working with the same dataset as the authors, excluding modern data for which information on several of these potentially confounding factors was not available, we analysed the effect of time on the three identifiable taxa. We grouped the data into three time periods

(ancient, medium and modern) and found a significant interaction: the effect of time varies across taxa (GLM, interaction taxon: time period, $\chi^2_{(3,54)} = 32.3$, $P < 0.0001$; Fig. 1b). Given that these taxonomic groups were not represented over the entire 300 million years considered (no medium time data for plants, almost exclusively medium time data for pollen), it is reasonable to consider to what extent the

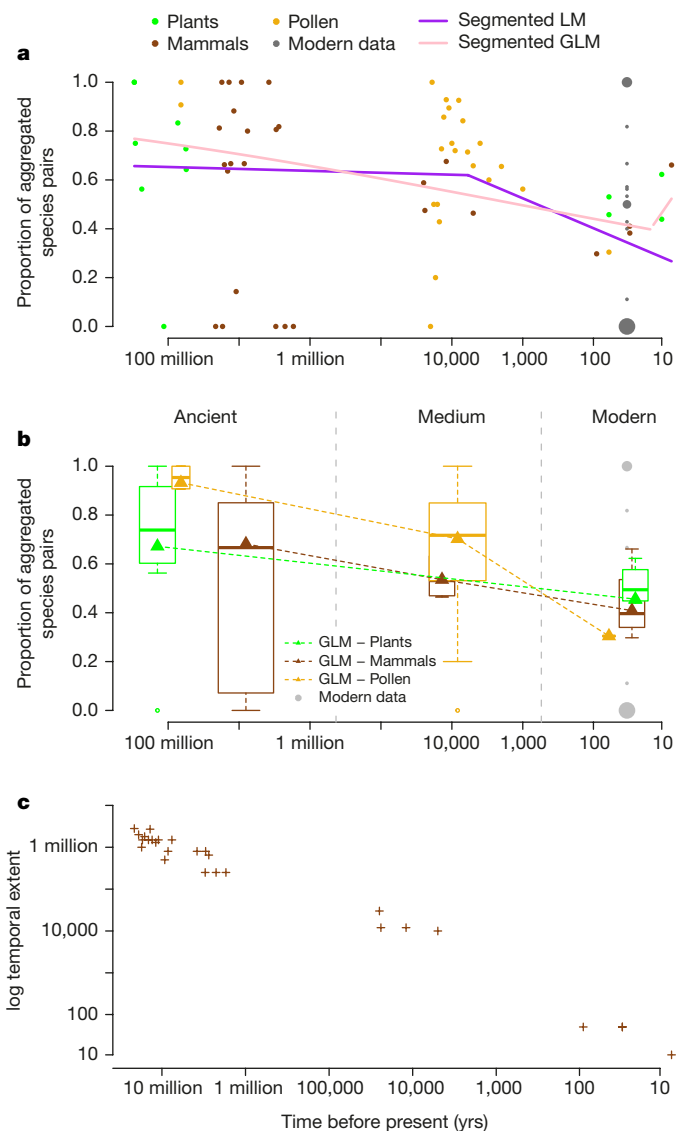


Figure 1 | Models of community structure over time: three types of possible bias. **a**, Proportion of aggregated species pairs over time, modelled using a linear segmented regression showing a breakpoint (model of Lyons *et al.*²) compared to a GLM with binomial error. LM, linear model. **b**, GLM of the proportion of aggregated species pairs by time period in interaction with the taxonomic group. Widths of the boxplots are proportional to the number of points per group and the boxplots are centred on the mean of the time period per group. **c**, Correlation of time before present and temporal extent for mammal data.

effect of taxonomic group, and the inclusion of modern data whose taxonomic groups we do not know, determine the detection of a breakpoint.

To test whether the proportion of aggregated species pairs declines over time when additional confounding factors are taken into account, we focused on mammals because they are the only group with at least four data points for each of the time periods (ancient, medium and modern). We found that ‘temporal extent’—the maximum amount of time encompassed by a dataset—was strongly correlated to ‘time before present’ ($R^2 = 0.98$, $F_{(1,25)} = 1,325$, $P < 0.0001$). It is impossible to determine which of the two variables (temporal resolution or time before present) causes the decrease in the proportion of aggregated species pairs. Therefore, the data of Lyons *et al.*² did not provide evidence for a decreasing proportion of aggregated species pairs over geological time scales. We propose as an alternative hypothesis that the confounding variable, temporal extent, may act on the proportion of aggregated species pairs in the following manner: by associating several communities over an increasingly long time window (up to 16 million years), the probability of obtaining positive associations by chance might increase. This could be the case if the same environmental features caused successions of different species to have similar geographical distributions although they never actually co-occurred.

The conclusions of Lyons *et al.*² rest on the assumption that natural communities of different taxonomic groups, spatial and temporal extents or geographical locations are comparable across time periods. We disagree and believe that none of their findings supports their claim that Holocene shifts in the assembly of plant and animal communities have occurred and implicate human impacts. We urge greater caution in conducting and interpreting community analyses over geological time scales. Robust generalizations about those shifts in driving forces of community structure will be possible only after incorporating a much higher number of assemblages, especially around the critical period encompassing a putative breakpoint. The study design should be based on the question that is addressed. If the authors aim to identify shifts in community structure due to human impacts, they should sample a greater number of assemblages in the Holocene and

Anthropocene³ when critical shifts in manmade environmental modifications occurred. It is unclear how ancient data (as old as 300 million years) would help to address this question. Crucially, future studies should tease apart impacts of confounding factors potentially influencing community structure.

Methods

Using a GLM (binomial error, logit link), we modelled the proportion of aggregated species pairs as a function of time and tested for a breakpoint using the R package ‘segmented’ (following ref. 2; Fig. 1a). We tested the interaction of time (treated as a factor with three groups) and taxonomic groups using a GLM without the ‘modern’ dataset (Fig. 1b). Using a GLM, we tested for a confounding effect of temporal extent and time. We modelled the proportion of aggregated species pairs as a function of temporal extent (Fig. 1c) and tested whether temporal extent and time were correlated (Pearson’s r).

Cleo Bertelsmeier¹ & Sébastien Ollier²

¹Université de Lausanne, Département d’écologie & évolution, Le Biophore, UNIL-Sorge, 1015 Lausanne, Switzerland.

email: cleo.bertelsmeier@unil.ch

²Ecologie Systématique Evolution, CNRS, Univ. Paris-Sud, AgroParisTech, Université Paris-Saclay, 91400 Orsay, France.

Received 1 February; accepted 4 July 2016.

1. Dietl, G. P. Ecology: Different worlds. *Nature* **529**, 29–30 (2016).
2. Lyons, S. K. *et al.* Holocene shifts in the assembly of plant and animal communities implicate human impacts. *Nature* **529**, 80–83 (2016).
3. Corlett, R. T. The Anthropocene concept in ecology and conservation. *Trends Ecol. Evol.* **30**, 36–41 (2015).

Supplementary Information is available in the online version of the paper.

Author Contributions C.B. and S.O. designed the study, analysed the data and wrote the manuscript.

Competing Financial Interests Declared none.

doi:10.1038/nature19110

Lyons *et al.* reply

REPLYING TO C. Bertelsmeier & S. Ollier *Nature* **537**, <http://dx.doi.org/10.1038/nature19110> (2016)

In the accompanying Comment¹, Bertelsmeier and Ollier criticize our statistical analyses and question our conclusion that there was a shift in community structure in the mid-Holocene epoch². The critique is based on a generalized linear model (GLM) using a binomial error distribution and logit link. We question the validity of the analyses made by Bertelsmeier and Ollier¹ for several reasons.

First, they suggest that our data² are better modelled by a GLM using a binomial distribution than by the normal distribution used in ordinary least squares (OLS) regression. However, in calculating model error, the GLM assumes a binomial error distribution for each counted Bernoulli event (here, species pairs in an assemblage), and that these are mutually independent. However, pairs are not independent of one another because the same species occurs in multiple pairs. The error distribution of species pairs across an assemblage is unknown, but it is not modelled by a simple binomial, in part because there are three possible outcomes (aggregated, segregated, or neither). Our linear model (LM) is simpler because it estimates the variance (error term) using deviations from the regression line and it fits the data better. By contrast, the binomial GLM makes strong assumptions about independence and error mean–variance

relationships. Using the same subset of the data, we found that Akaike information criterion (AIC) values provide much stronger support for a breakpoint analysis using OLS regression (74.6) than using a GLM (637.0).

Second, Bertelsmeier and Ollier¹ argue that datasets with only a few significant pairs should be excluded because those estimates are unreliable. They ignore the fact that significant pairs were identified using a null model that preserves row and column totals (species richness per site, and incidence per species) in the null assemblages. Aggregation must appear above and beyond the values expected from the number of species collected per sample. Finding limited significant aggregations is not an unreliable estimate, but a robust result of the null model.

Third, in re-testing for a Holocene shift, Bertelsmeier and Ollier¹ exclude most modern datasets, subset taxonomic groups, and break the data into three arbitrary time bins that bear no correspondence to accepted geological time periods³. Their ‘ancient’ is 300 million years ago to 1 million years ago, which encompasses two of the largest recorded mass extinctions, the break-up of Pangaea, considerable changes in climate, the rise of angiosperms and a change from

BRIEF COMMUNICATIONS ARISING

dinosaur- to mammal-dominated terrestrial ecosystems. Moreover, their definitions merge all Pleistocene and non-modern Holocene data into a single bin, obscuring a Holocene shift.

Using a two-way analysis of variance (ANOVA), they conclude that “the effect of time varies across taxa” and maintain that well-characterized plant and pollen assemblages should have been excluded because of their arbitrary time bins. Frankly, we are more impressed with similarities among the taxonomic subsets. All three subsets show a decrease in aggregations towards the modern; none shows a flat or increasing relationship, and pollen shows a mid-Holocene shift consistent with our previous conclusions².

Fourth, they propose that long temporal extents in older datasets increase aggregated pairs. We previously demonstrated that temporal extent cannot explain these results (see figure 2 in ref. 2). We further test this here by collapsing 32 time series into 8 time-averaged assemblages (North American mammals (3 datasets), Kenyan mammals (2 datasets), South African mammals (2 datasets), North American pollen: 0–7,000 years (8 datasets), 8,000–14,000 years (7 datasets), 15,000–20,000 (6 datasets) years and 65 million years ago (2 datasets), and Palaeocene–Eocene thermal maximum (PETM) plants (2 datasets)) with increased temporal extents and rerunning Pairs⁴. Of these, 22 showed a decrease in significant aggregations, 9 showed an increase and 1 showed no change. Time-averaging does not necessarily increase aggregations.

In summary, the critique of Bertelsmeier and Ollier¹ depends on (1) the use of a GLM for non-independent proportional data that do not follow a binomial distribution; (2) a fundamental misunderstanding of the fossil record and geological time; and (3) selective use of our data. Bertelsmeier and Ollier¹ regard our use of fossil datasets representing different taxonomic groups, spatial and temporal extents, and geographical locations as inappropriate for demonstrating a Holocene shift in community structure. We regard our diverse datasets as good support for the strength of the signal we detected, which emerged in spite of this diversity. Contrary to the assertion by Bertelsmeier and Ollier¹, finding a recent shift in an ancient pattern that has been stable for 300 million years is entirely relevant to disentangling the effects of humans and natural causes on community structure. We stand by our original analyses and conclusions.

The author order of this Reply reflects the relative contributions of the authors, and is different from that in ref. 2. Author D.W. did not participate in this response.

S. Kathleen Lyons¹, Joshua H. Miller², Anikó Tóth¹, Kathryn L. Amatangelo³, Anna K. Behrensmeyer¹, Antoine Bercovici¹, Jessica L. Blois⁴, Matt Davis⁵, William A. DiMichelle¹, Andrew Du⁶, Jussi T. Eronen⁷, J. Tyler Faith⁸, Gary R. Graves^{9,10}, Nathan Jud¹¹, Conrad Labandeira^{1,12,13}, Cindy V. Looy¹⁴, Brian McGill¹⁵, David Patterson⁶, Silvia Pineda-Munoz¹, Richard Potts¹⁶, Brett Riddle¹⁷, Rebecca Terry¹⁸, Werner Ulrich¹⁹, Amelia Villaseñor⁶, Scott Wing¹, Heidi Anderson²⁰, John Anderson²⁰ & Nicholas J. Gotelli²¹

¹Department of Paleobiology, National Museum of Natural History, Smithsonian Institution, Washington DC 20013, USA.
email: lyonss2@si.edu

²Department of Geology, University of Cincinnati, Cincinnati, Ohio 45221, USA.

³Department of Environmental Science and Biology, The College at Brockport – SUNY, Brockport, New York 14420, USA.

⁴School of Natural Sciences, University of California, Merced, 5200 North Lake Road, Merced, California 95343, USA.

⁵Department of Geology and Geophysics, Yale University, New Haven, Connecticut 06520, USA.

⁶Hominid Paleobiology Doctoral Program, Center for the Advanced Study of Hominid Paleobiology, Department of Anthropology, George Washington University, Washington DC 20052, USA.

⁷Department of Geosciences and Geography, University of Helsinki, 00014 Helsinki, Finland.

⁸School of Social Science, The University of Queensland, Brisbane, Queensland 4072, Australia.

⁹Department of Vertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington DC 20013, USA.

¹⁰Center for Macroecology, Evolution and Climate, University of Copenhagen, Copenhagen 2100, Denmark.

¹¹Florida Museum of Natural History, University of Florida, Gainesville, Florida 32611, USA.

¹²Department of Entomology, University of Maryland College Park, College Park, Maryland 20742, USA.

¹³Key Lab of Insect Evolution and Environmental Changes, Capital Normal University, Beijing 100048, China.

¹⁴Department of Integrative Biology and Museum of Paleontology, University of California Berkeley, Berkeley, California 94720, USA.

¹⁵School Biology and Ecology & Sustainability Solutions Initiative, University of Maine, Orono, Maine 04469, USA.

¹⁶Department of Anthropology, Human Origins Program, National Museum of Natural History, Smithsonian Institution, Washington DC 20013, USA.

¹⁷School of Life Sciences, University of Nevada-Las Vegas, Las Vegas, Nevada 89154, USA.

¹⁸Department of Integrative Biology, Oregon State University, Corvallis, Oregon 97731, USA.

¹⁹Chair of Ecology and Biogeography, Nicolaus Copernicus University, Lwowska 1, 87-100 Torun, Poland.

²⁰Evolutionary Studies Institute, University of the Witwatersrand, Johannesburg, South Africa.

²¹Department of Biology, University of Vermont, Burlington, Vermont 05405, USA.

1. Bertelsmeier, C. & Ollier, S. Questioning Holocene community shifts. *Nature* **537**, <http://dx.doi.org/10.1038/nature19110> (2016).
2. Lyons, S. K. *et al.* Holocene shifts in the assembly of plant and animal communities implicate human impacts. *Nature* **529**, 80–83 (2016).
3. Gradstein, F. M. *The Geologic Time Scale 2012* 1st edn (Elsevier, 2012).
4. Ulrich, W. Pairs—a FORTRAN program for studying pair wise species associations in ecological matrices. Available at: <http://www.keib.umk.pl/pairs/?lang=en> (2010).

doi:10.1038/nature19111

ALZHEIMER'S DISEASE

Attack on amyloid- β protein

An antibody therapy markedly reduces aggregates of amyloid- β , the hallmark protein of Alzheimer's disease, and might slow cognitive decline in patients. Confirmation of a cognitive benefit would be a game-changer. [SEE ARTICLE P.50](#)

ERIC M. REIMAN

It is 25 years¹ since the amyloid- β (A β) protein was proposed as the trigger for a cascade of events in the brain that lead to Alzheimer's disease. A growing number of anti-A β treatments have been developed to short-circuit this cascade — and several are currently being evaluated in people who have already developed or are at risk of developing symptoms of Alzheimer's (www.alzforum.org/therapeutics). On page 50, Sevigny *et al.*² report findings from an initial 12-month, placebo-controlled trial of the antibody aducanumab, which selectively binds to potentially harmful soluble and insoluble A β aggregates, respectively called A β oligomers and fibrils.

The trial was primarily intended to clarify the A β -fibril-reducing effects and safety of different aducanumab doses administered intravenously once a month. It involved people who had been diagnosed with mild cognitive impairment (non-disabling memory and thinking problems) or mild dementia (which did have a slightly disabling effect) due to Alzheimer's disease. Each of the participants also tested positive for A β in a positron emission tomography (PET) scan, indicating moderate to frequent build-up of fibril-containing plaques — a cardinal feature of the disease. The study was not designed to definitively address

aducanumab's effect on cognitive decline.

Aducanumab treatment was associated with unusually striking, progressive, dose-dependent reductions in PET measurements of A β -plaque burden. Aducanumab was also presumed to bind to and remove harder-to-measure A β oligomers, which seem to accumulate at or near plaques and may be the more damaging of the two aggregates³. What's more, despite the relatively small number of study participants and the substantial extent to which the disease has progressed by the time people with Alzheimer's develop memory and thinking problems, exploratory analyses suggested that higher antibody doses and greater A β -plaque reductions were associated with slower cognitive decline. If these preliminary cognitive findings are confirmed in larger and more-definitive clinical trials, which are now under way, it would provide a shot in the arm in the fight against Alzheimer's disease and compelling support for the amyloid hypothesis.

The amyloid hypothesis contends that a 42-amino-acid form of A β (A β ₄₂) becomes harmful when, owing to its overproduction or reduced clearance from the brain, individual A β ₄₂ monomers come together in various numbers and conformations to form oligomers and fibrils. These A β ₄₂ aggregates trigger a cascade of neurobiological events, including:

certain inflammatory responses; aggregation, phosphorylation and propagation of a protein called tau; and other neuronal changes. These events contribute to the formation of A β plaques and tau-containing tangles, loss of neurons and the synaptic connections between them, cognitive decline and disability, and other features of Alzheimer's disease (Fig. 1).

Proponents of the amyloid hypothesis cite an abundance of supporting evidence¹. Others note that the evidence is largely circumstantial, and that questions remain about the offending A β species and its effects. As such, they wonder whether A β ₄₂ accumulation is a consequence rather than an initiator of disease, and worry that anti-A β drug development might lead to a dead end. What will it take to confirm or refute the amyloid hypothesis once and for all?

Confirmation of this hypothesis will require definitive evidence that an anti-A β treatment can reduce cognitive decline in people affected by or at risk of developing Alzheimer's disease. Sevigny and colleagues' trial provides convincing evidence that aducanumab can enter the brain, target A β fibrils and substantially reverse plaque deposition — a major advance. But although the authors' additional cognitive findings are encouraging, they are not definitive. It would be prudent to withhold judgement about aducanumab's cognitive benefit until results from the larger trials are in. It will

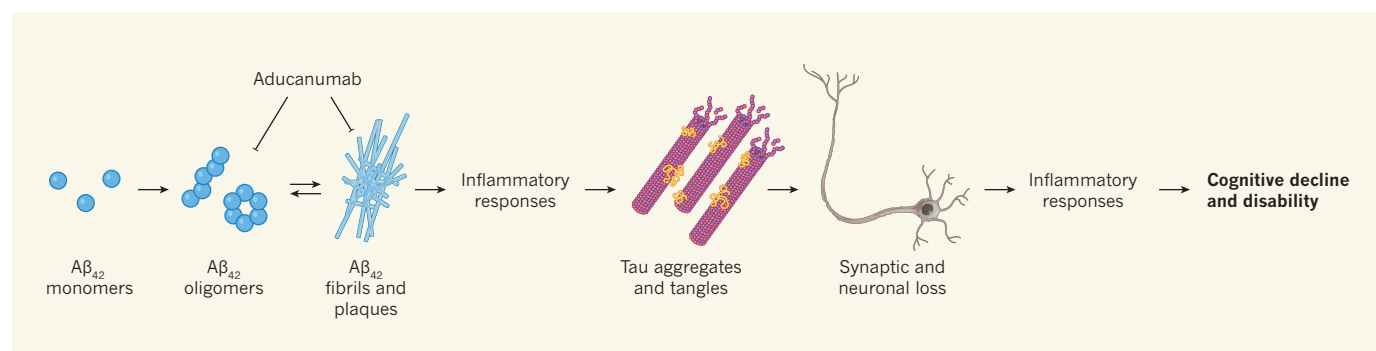


Figure 1 | The amyloid hypothesis. This hypothesis contends that increases in the amyloid- β 42 (A β ₄₂) protein trigger a cascade of events in the brain that lead to Alzheimer's disease. Under this hypothesis, individual A β ₄₂ monomers aggregate into damaging oligomers and fibrils or near A β ₄₂ plaques. A β aggregates cause certain inflammatory responses. Through unknown mechanisms, these events lead to the aggregation, phosphorylation and propagation of tau, a protein that is associated with microtubules (pink

and purple) and is the main constituent of harmful tangles. Affected neurons and synapses become dysfunctional and can die, leading to additional inflammatory responses. The progressive dysfunction, degeneration and loss of affected neurons and synapses is associated with cognitive decline, other symptoms of Alzheimer's disease and increasing disability. Sevigny *et al.*² report that the A β -binding antibody aducanumab binds to, promotes removal and blocks accumulation of fibrils and oligomers.

also be useful to see what can be learnt from large trials of other anti-A β treatments in the coming months and years.

Conversely, refutation of the amyloid hypothesis will require failure of anti-A β treatments to reduce cognitive decline in sufficiently large and suitably designed trials — not only in people with cognitive impairment due to Alzheimer's disease, but also in people without such impairment who have evidence of A β plaques, and even people without impairment who are at genetic risk of developing Alzheimer's but have little or no A β deposition. Several prevention trials using anti-A β treatments have started⁴, and more are on the way. Because abnormal A β build-up can begin more than two decades before the onset of memory and thinking problems⁵, having a drug such as aducanumab that substantially reverses pre-existing A β deposition might increase the chances of extinguishing the disease even after it has set in.

What accounts for aducanumab's unusually pronounced plaque-busting effects, even in small doses and despite the fact that only one to two antibody molecules out of every thousand are thought⁶ to cross the blood–brain barrier? It might be a combination of three things: the drug's unusually high selectivity for A β ₄₂ fibrils and oligomers, which minimizes the number of antibody molecules that bind to the abundant A β monomers in the blood and so maximizes the number of unbound antibodies that can enter the brain; its unusually high affinity for A β ₄₂ fibrils and oligomers; and the mechanism by which it enlists microglia, the brain's principal immune cells, to engulf and clear A β fibrils.

On the one hand, aducanumab's microglia-mediated activity could account for the antibody's ability to remove plaques, rather than just to slow further A β accumulation (which would be valuable in its own right). On the other hand, this activity might increase the chance of people developing amyloid-related imaging abnormalities (ARIA) — defects characterized by evidence of brain-fluid accumulation in magnetic resonance imaging scans. Like certain other anti-A β antibody treatments⁷, Sevigny and colleagues' study found that aducanumab was more likely to cause ARIA in higher doses and in people who carry the *APOE* type 4 gene, which is the major genetic risk factor for Alzheimer's disease.

The authors observed that ARIA were sometimes associated with transient headaches, visual disturbances or confusion, but were often associated with no symptoms, and that symptoms typically resolved within one to three months. Nonetheless, the frequency of ARIA caused the researchers to limit the maximum dose studied. It will be important to establish a sweet spot: a dose that is sufficiently safe and well tolerated, but also effective.

In addition to confirming the amyloid

hypothesis, finding that the effects of treatments such as aducanumab on A β or other biological measurements of Alzheimer's disease are associated with a cognitive benefit might help to accelerate the evaluation and regulatory approval of promising Alzheimer's prevention therapies that are based on reducing the biological measurements alone⁴. Indeed, confirmation that an anti-A β treatment slows cognitive decline would be a game-changer for how we understand, treat and prevent Alzheimer's disease. Now is the time to find out. ■

Eric M. Reiman is at the Banner Alzheimer's

Institute, Phoenix, Arizona 85006 USA.
e-mail: eric.reiman@bannerhealth.com

1. Selkoe, D. J. & Hardy, J. *EMBO Mol. Med.* **8**, 595–608 (2016).
2. Sevigny, J. *et al.* *Nature* **537**, 50–56 (2016).
3. Haass, C. & Selkoe, D. J. *Nature Rev. Mol. Cell Biol.* **8**, 101–112 (2007).
4. Reiman, E. M. *et al.* *Nature Rev. Neurol.* **12**, 56–61 (2016).
5. Fleisher, A. S. *et al.* *JAMA Neurol.* **72**, 316–324 (2015).
6. Yu, J. Y. & Watts, R. J. *Neurotherapeutics* **10**, 459–472 (2013).
7. Sperling, R. A. *et al.* *Alzheimer's Dement.* **7**, 367–385 (2011).

The author declares competing financial interests. See online article for details.

PLANETARY SCIENCE

Cometary dust under the microscope

The Rosetta spacecraft made history by successfully orbiting a comet. Data from the craft now reveal the structure of the comet's dust particles, shedding light on the processes that form planetary systems. [SEE LETTER P.73](#)

LUDMILLA KOLOKOLOVA

Planetary systems such as the Solar System were built from dust in protoplanetary nebulae, the clouds of gas and dust in which stars and planets are born. These dust particles collided, stuck together and eventually formed planetesimals, the building blocks of planets. Comets are leftover planetesimals, made of ice and dust, and range from hundreds of metres to tens of kilometres in diameter. They spend most of their lives on the outskirts of the Solar System — away from damaging radiation and high temperatures, and avoiding collisions with other objects — thus preserving the material that originally formed the protoplanetary nebula. By studying comets, we can learn about the processes that gave rise to the Solar System, even though those processes happened almost five billion years ago¹. On page 73, Bentley *et al.*² show that cometary dust particles are formed from a hierarchical assembly of smaller constituents, a discovery that has implications for our understanding of the formation and evolution of planetary systems.

Because we cannot catch a comet and study it in the laboratory, previous analyses have inferred the properties of cometary dust particles from their interactions with sunlight. One of the earliest such analyses³ indicated that these particles are not solid, compact objects, but loosely packed aggregates of tiny (sub-micrometre diameter) grains. An aggregate structure was also found in interplanetary

dust particles (IDPs) collected in Earth's upper atmosphere. Many of these IDPs were found to have originated from comets⁴.

More evidence for the aggregate nature of cometary dust came from the Stardust spacecraft, which collected dust particles during its close fly-by⁵ of the comet Wild 2. However, neither the IDPs nor the Stardust samples were unmodified (pristine). The IDPs would have been affected by their long exposure to solar radiation, any collisions with other dust particles and interactions with Earth's atmosphere. In the case of the Stardust samples, the spacecraft collected dust particles at a distance of hundreds and even thousands of kilometres from the comet's surface — and as the particles travelled between the two, their properties would have changed as a result of evaporation of volatile components, possible destruction of complex organic compounds and fragmentation of the particles themselves.

The IDPs and Stardust samples were also damaged, or even completely shattered, during collection. In the case of the Stardust samples, because the dust particles were travelling at a speed of 6.1 kilometres per second relative to the spacecraft, the particles were damaged by the impact with collecting cells in the Stardust sample collector. As a result, the aggregate structure of the particles was not measured directly. Instead, the structure was inferred from the complex shape of the tracks that the particles produced while crossing the aerogel — a low-density material — in the cells, or from the impact craters they left on

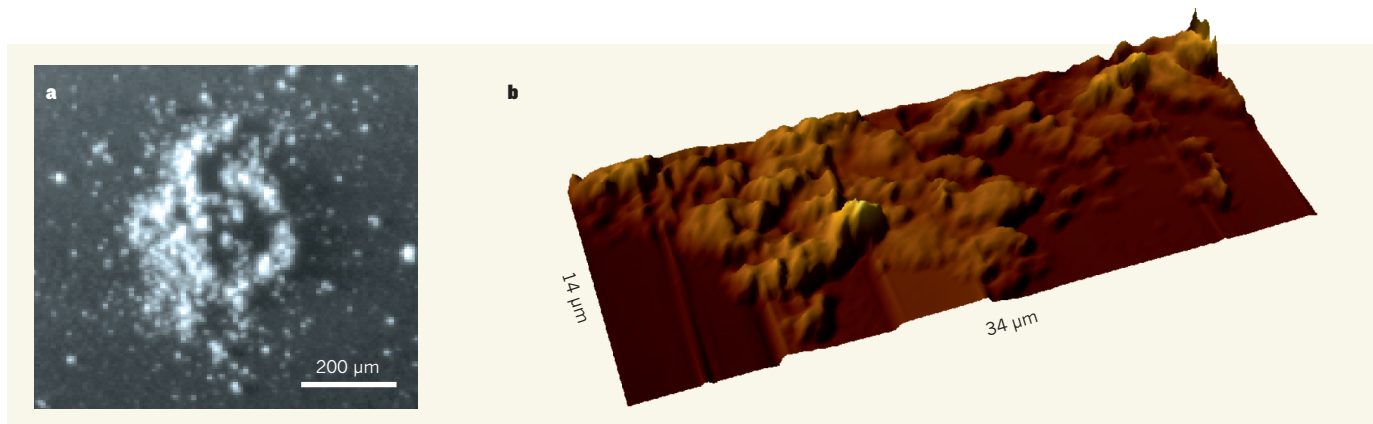


Figure 1 | Dust from comet 67P/Churyumov–Gerasimenko. **a**, A dust particle collected by the Cometary Secondary Ion Mass Analyser (COSIMA)⁷. The image shows that cometary dust particles are aggregates of grains with diameters larger than a few micrometres. **b**, Bentley *et al.*² use an atomic force microscope to measure the size, shape and texture of cometary dust with greater resolution than COSIMA. The 3D image shows an example of one of the dust particles analysed. The authors' results indicate that the COSIMA grains are built from even smaller, sub-micrometre grains. When combined with the data from COSIMA, Bentley and colleagues' findings show that cometary dust particles are created by the hierarchical assembly of smaller constituents.

ESA/ROSETTA/MPS FOR COSIMA TEAM/LANGEVIN ET AL. (2016)

the aluminium foil that separated the aerogel cells⁶. These limitations in data collection and analysis left the cometary dust particles' pristine structure undetermined.

A unique opportunity to study pristine dust particles arose when the Rosetta spacecraft came within tens of kilometres of comet 67P/Churyumov–Gerasimenko and obtained samples of cometary dust from this distance. These dust particles were first analysed using an instrument called the Cometary Secondary Ion Mass Analyser (COSIMA), which collected and imaged aggregate particles hundreds of micrometres in diameter (Fig. 1a). The images from COSIMA revealed that the particles were aggregates of grains with diameters larger than a few micrometres⁷, contradicting the earlier studies of IDPs and Stardust samples that indicated sub-micrometre grains.

Bentley and colleagues resolve this contradiction using data from Rosetta's Micro-Imaging Dust Analysis System (MIDAS). MIDAS is an atomic force microscope — it scans the collected dust particles using a sharp, needle-like tip, which provides a 3D image of the particles with a maximum resolution of about 4 nm (Fig. 1b). The images from MIDAS show that the grains seen by COSIMA are built from even smaller, sub-micrometre grains. The authors' discovery not only proves that the basic building blocks of cometary dust particles are sub-micrometre grains, but also reveals the hierarchical nature of dust particles.

This hierarchical structure, although not detected by remote-sensing studies or analyses of the Stardust samples, had been hypothesized by some researchers. For example, models of the upper layers of cometary surfaces provided the most realistic results when these layers were assumed to consist of hierarchically structured dust particles⁸. Another analysis⁹ found that hierarchical growth is necessary to reproduce the dust-size distribution that provides the best fit to characteristics of observed protoplanetary nebulae.

In addition to showing that cometary dust

particles have a hierarchical structure, Bentley *et al.* find that the basic building blocks of the particles are roughly spheroidal — the shape of a deformed (elongated or flattened) sphere. By approximating the grains as spheroids, the authors find that their large axis is, on average, 2.87 times longer than the small axis. These findings are strikingly similar to a model proposed by astrophysicist Mayo Greenberg¹⁰ in the 1980s. In Greenberg's model, cometary dust particles are aggregates of interstellar dust particles, which are described as spheroids with an axis ratio of 3 to 1.

The authors' results enhance our fundamental understanding of cometary dust, and the processes that ultimately gave rise to planetary systems such as the Solar System. Their discovery of a hierarchical structure in cometary dust particles and their description of the basic building blocks of such particles might lead physicists to reconsider the interpretation of data obtained from ground-based observations of comets and re-evaluate the processes in protoplanetary nebulae — and will

probably give rise to new models of how planets were formed. ■

Ludmilla Kolokolova is in the Department of Astronomy, University of Maryland, College Park, Maryland 20742, USA.
e-mail: ludmilla@astro.umd.edu

1. Tilton, G. R. in *Meteorites and the Early Solar System* (eds Kerridge, J. F. & Matthews, M. S.) 259–275 (Univ. Arizona Press, 1988).
2. Bentley, M. S. *et al. Nature* **537**, 73–75 (2016).
3. Weiss-Wrana, K., Giese, R. H. & Zerull, R. H. in *Properties and Interactions of Interplanetary Dust* (eds Giese, R. H. & Lamy, P.) 223–226 (Springer, 1985).
4. Sandford, S. A. *Fundament. Cosmic Phys.* **12**, 1–73 (1987).
5. A'Hearn, M. F. *Nature* **429**, 818–819 (2004).
6. Kearsley, A. T. *et al. Meteor. Planet. Sci.* **43**, 41–73 (2008).
7. Schulz, R. *et al. Nature* **518**, 216–218 (2015).
8. Skorov, Y. & Blum, J. *Icarus* **221**, 1–11 (2012).
9. Dominik, C. in *Vol. 414: Cosmic Dust — Near and Far* (eds Henning, T., Grün, E. & Steinacker, J.) 494 (Astron. Soc. Pacif., 2009).
10. Greenberg, J. M. in *Asteroids, Comets, Meteors II* (eds Lagerkvist, C.-I., Rickman, H., Lindblad, B. A. & Lundstedt, H.) 221–223 (Uppsala Univ., 1986).

STRUCTURAL BIOLOGY

Moulding the ribosome

Production of the cell's translational apparatus, the ribosome, requires the orchestrated function of hundreds of proteins. A structure of its earliest precursor yields unprecedented insight into ribosome formation.

MARLENE OEFFINGER

In every living cell, a large macromolecular complex called the ribosome is responsible for translating messenger RNA into amino-acid chains in the cytoplasm. A mature ribosome contains about 80 ribosomal proteins (r-proteins) and four ribosomal RNAs (rRNAs). Yet the construction of a ribosome

is mediated by many more proteins and RNA molecules within large dynamic pre-ribosomal complexes. Writing in *Cell*, Kornprobst *et al.*¹ report that they have exploited advances in cryo-electron microscopy² to resolve the structure of the earliest pre-ribosome, the 90S, to a near-atomic resolution of between 4 and 7 ångströms. The structure reveals, for the first time and in stunning detail, the arrangement of

and interactions between many proteins that have been implicated in ribosome assembly, shedding light on a crucial step in early ribosome formation.

In 1967, it was discovered³ that, in eukaryotic organisms (those whose cells carry a nucleus), a long RNA transcript called the pre-rRNA undergoes processing in a nuclear compartment, the nucleolus, to produce three of the four rRNAs found in the mature ribosome. An analysis⁴ later that year of ribosomes isolated from human nuclei, and a comparison⁵ of cytoplasmic and nuclear ribosomes in 1972, revealed that nuclear ribosomes contain many more proteins than do their cytoplasmic counterparts. These extra proteins were hypothesized to help process the pre-rRNA.

Since then, the steps of pre-rRNA processing have been established and most of the extra proteins (now called ribosome biogenesis factors) have been identified, thanks to advances in biochemistry and mass spectrometry. During its transcription, the long pre-rRNA is assembled with r-proteins, ribosome biogenesis factors and small nucleolar RNAs to form a large 90S pre-ribosome. Following the first stage of pre-rRNA processing, the complex splits into two pre-ribosomes, dubbed pre-40S and pre-60S, which are eventually exported to the cytoplasm where they undergo further maturation steps and then join as 40S and 60S subunits to form the mature ribosome.

Along with the identities of the biogenesis factors came the realization that they numbered a vast 200 to 300 in eukaryotes^{6,7}. In the yeast *Saccharomyces cerevisiae*, the 90S pre-ribosome alone contains about 70 ribosome biogenesis factors — almost as many as the number of proteins in a mature ribosome⁶. Hence, a recurring question in the field is: why does ribosome production require so many accessory proteins?

By resolving the structure of the 90S pre-ribosome in the yeast *Chaetomium thermophilum*, Kornprobst *et al.* provide an answer to this question. The authors identified features in their structure by fitting data from previous biochemical and genetic studies (including X-ray structures of several proteins, predicted protein-domain structures and known protein–protein and protein–pre-rRNA interactions) to determine where different proteins and RNAs are located in the 90S complex. The requirement for so many extra proteins is explained by the authors' observation that many accessory proteins are arranged around the folded pre-rRNA molecule in previously defined⁸ multi-protein complexes called UTP-A, UTP-B and UTP-C. Of these, UTP-A and UTP-B form a scaffold, within which the newly transcribed pre-rRNA is encased and so can be securely processed, modified and assembled with r-proteins (Fig. 1).

The role of this scaffold is reminiscent of the way in which chaperone proteins aid folding of other proteins — a common process

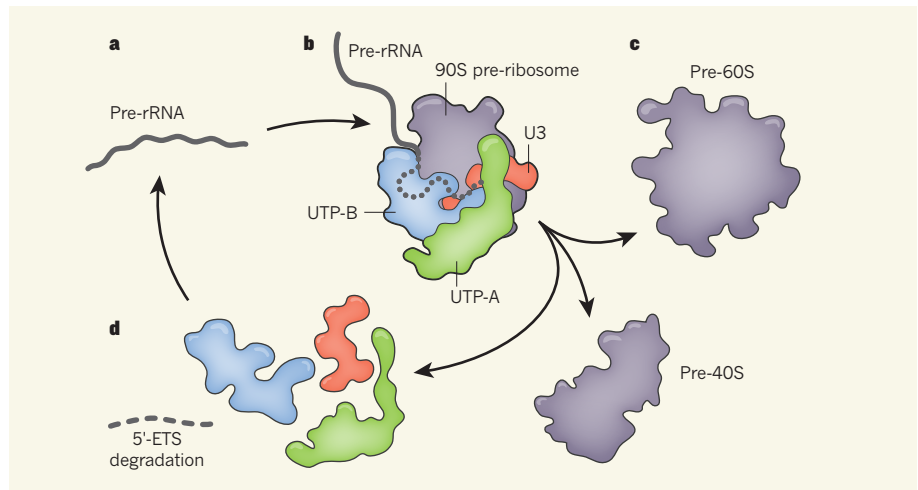


Figure 1 | Maturation of the 90S pre-ribosome. Ribosome assembly involves processing of a long RNA transcript called the pre-rRNA by a complex known as the 90S pre-ribosome. Kornprobst *et al.*¹ resolved the structure of the 90S in near-atomic resolution. **a**, A pre-rRNA is generated. **b**, The authors show that the pre-rRNA is threaded into a mould formed by the protein complexes UTP-A and UTP-B, and the RNA–protein complex U3. Encased within this mould, the pre-rRNA is safely folded and processed, with a sequence called the 5' external transcribed spacer (5'-ETS) being cleaved away (hidden from view). **c**, After processing, pre-40S and pre-60S complexes, which will go on to form the ribosome, separate from the mould components of the 90S. **d**, The UTP and U3 complexes are presumably recycled for use with the next pre-RNA, whereas the excised 5'-ETS is degraded.

that prevents aggregation of proteins into non-functional structures. But although chaperone-mediated protein folding has been long established⁹, the idea of chaperone moulds is new to RNA biology.

The 90S chaperone mould also includes the small nucleolar ribonucleoprotein complex U3 — an RNA–protein complex that has known roles in pre-rRNA processing and folding^{10,11}. Kornprobst *et al.* showed that one half of U3 spans the outer body of the 90S complex in a scaffold-like arrangement, whereas the other half is buried deep within the 90S, presumably interacting with the pre-rRNA. This part of U3 is associated with a region at the end of the pre-rRNA called the 5' external transcribed spacer (5'-ETS), and the authors demonstrated that cleavage of this spacer from the pre-rRNA is crucial for the separation of the processed 90S pre-rRNAs into pre-40S and pre-60S complexes, and the progression of ribosome production.

Kornprobst and colleagues also identified the position of the pre-18S rRNA (which will become the rRNA component of the 40S subunit) in their structure. When comparing the pre-18S structure with that of the mature 18S rRNA, the authors observed that the molecule underwent progressive folding, beginning in the domains closest to the site where transcription began. In the 90S, these regions were folded to resemble the mature 18S, whereas domains farther from the transcriptional start site were seemingly still in transitory states. This observation fits well with a previous model⁶ of hierarchical rRNA assembly.

Kornprobst and colleagues have visualized in detail what, until now, has been seen

through electron microscopy only as small black balls on strings of pre-rRNA. Holding a magnifying glass to the early steps of ribosome biogenesis, the authors have finally revealed a role for the multitude of ribosome biogenesis factors as a chaperone mould that provides a secure environment for the processing and folding of pre-rRNA.

The 90S pre-ribosome contains the entire rRNA precursor, which includes several transcribed spacer sequences that will be cleaved away, and sequences that will give rise to the rRNAs of the 60S ribosomal subunit. However, Kornprobst *et al.* focused on only the rRNA region and the proteins that give rise to the 40S subunit. As such, many questions about 60S formation remain unanswered — for instance, whether a separate chaperone-like mould encases these other regions of the pre-rRNA.

There are several structures visible in 90S that have not yet been identified. In years to come, it will be interesting to index these features and further unravel the role of the UTP-C complex and other proteins in 90S pre-rRNA maturation. Using the technical advances highlighted in the current study, we can hope to shed more light on the dynamic and multi-tiered process that is ribosome formation. ■

Marlene Oeffinger is in the Department for Systems Biology, Institut de recherches cliniques de Montréal, Montreal, Quebec H2W 1R7, Canada.
e-mail: marlene.oeffinger@ircm.qc.ca

1. Kornprobst, M. *et al.* *Cell* **166**, 380–393 (2016).
2. Kühlbrandt, W. *eLife* **3**, e03678 (2014).
3. Weinberg, R. A., Loening, U., Willems, M. & Penman, S.

- Proc. Natl Acad. Sci. USA* **58**, 1088–1095 (1967).
 4. Warner, J. R. & Soeiro, R. *Proc. Natl Acad. Sci. USA* **58**, 1984–1990 (1967).
 5. Kumar, A. & Warner, J. R. *J. Mol. Biol.* **63**, 233–246 (1972).
 6. Woolford, J. L. & Baserga, S. J. *Genetics* **195**,

- 643–681 (2013).
 7. Henras, A. K., Plisson-Chastang, C., O'Donohue, M.-F., Chakraborty, A. & Gleizes, P.-E. *Wiley Interdisc. Rev. RNA* **6**, 225–242 (2014).
 8. Chaker-Margot, M., Hunziker, M., Barandun, J., Dill, B. D. & Klinge, S. *Nature Struct. Mol. Biol.* **22**,

- 920–923 (2015).
 9. Hartl, F. U., Bracher, A. & Hayer-Hartl, M. *Nature* **475**, 324–332 (2011).
 10. Dragon, F. *et al. Nature* **417**, 967–970 (2002).
 11. Dutca, L. M., Gallagher, J. E. G. & Baserga, S. J. *Nucleic Acids Res.* **39**, 5164–5180 (2011).

CLIMATE SCIENCE

Southern Ocean freshened by sea ice

The Southern Ocean has become less salty during the past few decades. An analysis of sea-ice transport in the ocean suggests that this phenomenon can be explained by coupled changes in sea-ice drift and thickness. [SEE LETTER P.89](#)

TED MAKSYM

The vast band of water that encircles the Antarctic continent, known as the Southern Ocean, is the world's dominant ocean sink for heat and carbon dioxide¹. It also has a crucial role in the global overturning circulation — the sinking, at high latitudes, of cold, dense surface waters to the deep ocean, and the compensatory rising of deep waters originating from lower latitudes. The Southern Ocean's salinity has fallen during the past half-century, in the surface and intermediate waters of the open ocean^{2,3} and coastal regions⁴, and in deeper waters⁵. This freshening of surface waters has increased stratification (the vertical gradient of water density), potentially inhibiting upwelling of deeper water and affecting CO₂ uptake⁶. On page 89, Haumann *et al.*⁷

show that the freshening can be explained by changes in Antarctic sea-ice production and transport (Fig. 1).

Previous explanations for this freshening have included a net increase in the difference between the amount of precipitation and the amount of evaporation over the ocean⁸, and increased input of glacial meltwater^{4,5}. However, the former is inadequate for explaining the freshening in surface and intermediate waters of the open ocean, and the latter overestimates the freshening of deep waters. The constant movement of sea ice redistributes a substantial amount of fresh water⁹, so Haumann and colleagues chose to investigate the potential contribution of sea-ice transport to this observed freshening.

When sea ice forms, most of the salt is lost to the upper ocean, so the ocean loses fresh water

and its salinity increases. When the ice melts, the fresh water is returned to the ocean. The net impact on the upper ocean would be minimal were it not for prevailing winds that tend to push the ice from coastal waters, where most of it is formed, to the north, where it melts. This drives a net transport of fresh water that contributes to the overturning circulation of the Southern Ocean. The saltier water that results from ice formation in coastal regions contributes to the generation of Antarctic Bottom Water, and the fresh meltwater input to the north mixes with upwelling deep water to modify the upper waters of the open Southern Ocean⁹.

Determining any trends in this sea-ice-driven freshwater transport is challenging, in part because of a lack of reliable data. The volume of sea ice transported can be calculated as the product of ice concentration (the fractional area of the ocean covered by ice), ice thickness and ice drift rate. However, satellite-derived ice-drift rates have significant biases relative to those measured by drifting buoys, and potential biases due to changes in data sources and satellite sensors over time. And there is no long-term data set for ice thickness.

Haumann *et al.* addressed these challenges by carefully reconstructing time series for each of these variables for the period from 1982 to 2008. First, they established a consistent satellite ice-drift time series by removing inconsistent data associated with the



WOLFGANG KAEHLER/LIGHTROCKET/GETTY

Figure 1 | Sea ice in the Southern Ocean. Haumann *et al.*⁷ report that changes in Antarctic sea-ice drift have altered the salinity of the Southern Ocean.

transitions between satellites, and stitched together different time periods by correcting for estimated biases. They then scaled the satellite ice-drift series to make it consistent with observed buoy drift.

To reconstruct a time series of ice thickness, the authors turned to a model-based estimate of ice-thickness trends constrained by observations of ice concentration. They then adjusted for potential biases in the modelled thicknesses using both sparse *in situ* data¹⁰ and ice-thickness estimates from satellite data¹¹. The time series for ice drift and thickness allowed Haumann *et al.* to make more-robust estimates of freshwater transport than were previously possible. This, in turn, allowed them to estimate the impact of transport trends on the salinity of the Southern Ocean using a simple model of water-mass exchange between the surface and the deeper waters.

The researchers show that the net transport of sea-ice-driven fresh water is substantial: larger than the inputs from glacial melt and comparable to the net input of precipitation and evaporation⁹. The estimated temporal trends are also sizeable: there is a 20% increase in transport over the 26-year study period. Notably, however, there is considerable regional variability in freshwater transport trends, including a large increase in the Pacific sector of the Southern Ocean (which encompasses the Ross Sea, where positive trends in northward ice drift and extent are largest¹²). Transport has decreased slightly elsewhere. Overall, Haumann *et al.* estimate that sea-ice-driven transport has contributed enough fresh water to the open-ocean surface and intermediate waters to explain the observed freshening.

A compelling result is that the calculated trends in sea-ice-driven freshwater transport are consistent with other observed patterns of change. First, the increases in freshwater transport occur in the Pacific sector, where increased freshening in surface waters has been strongest². Second, the increase in salt input due to sea-ice production in the coastal Pacific sector might explain why the observed freshening of Antarctic Bottom Water is less than that predicted from increased glacial melt⁵.

It is striking that major changes to ocean properties can occur as a result of relatively small average changes in sea-ice cover. Sea-ice extent has increased only slightly overall during the period covered by the time series, albeit with strongly contrasting regional patterns of change¹³. These regional changes were partly wind-driven¹², but, as Haumann *et al.* show, there may be little to no trend in the mean drift speed of sea ice. This demonstrates that it is the coupled trends in regional ice thickness and ice drift that are key to driving freshwater redistribution.

An important caveat to the findings is that the uncertainty in the derived trends is

considerable, and potentially underestimated. The corrections for bias in ice drift are large, and are difficult to quantify for the earlier years, for which there are almost no independent data available to provide validation. Nevertheless, the authors' estimates of freshwater transport remain similar when they are based on ice drift estimated from surface winds, which are a reasonable proxy for drift. The need for better ice-thickness estimates is also clear; ice thickness is the largest source of uncertainty in the results, and ice-thickness trends are the least well constrained by observations. However, a recent complementary study⁹ that used a broader array of observations collected between 2005 and 2010 to constrain a coupled ice-ocean model broadly supports the regional patterns of sea-ice-driven freshwater transport estimated in the current study, allaying concerns about the uncertainties.

Haumann and colleagues' findings emphasize that Antarctic sea ice is not merely a passive indicator of climate change and variability, but also a driver of changes in the climate system. Through its potential influence on ocean stratification and CO₂ uptake, sea ice might have a bigger role than previously thought.

The implications of these results for the Southern Ocean in a warming world are uncertain, because climate models do not properly capture the observed changes in Antarctic sea ice¹⁴. However, anticipated future declines in ice extent and volume would suggest that sea-ice freshwater transport should decrease. If so, then future losses of sea ice can be expected to play a prominent part in changes in the Southern Ocean's overturning circulation. ■

Ted Maksym is in the Applied Ocean Physics and Engineering Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543-1050, USA.
e-mail: tmaksym@whoi.edu

- Frölicher, T. L. *et al.* *J. Clim.* **28**, 862–886 (2015).
- Wong, A. P. S., Bindoff, N. L. & Church, J. A. *Nature* **400**, 440–443 (1999).
- Böning, C. W., Dispert, A., Visbeck, M., Rintoul, S. R. & Schwarzkopf, F. U. *Nature Geosci.* **1**, 864–869 (2008).
- Jacobs, S. S. & Giulivi, C. F. *J. Clim.* **23**, 4508–4524 (2010).
- Purkey, S. G. & Johnson, G. C. *J. Clim.* **26**, 6105–6122 (2013).
- de Lavergne, C., Palter, J. B., Galbraith, E. D., Bernardello, R. & Marinov, I. *Nature Clim. Change* **4**, 278–282 (2014).
- Haumann, F. A., Gruber, N., Münnich, M., Frenger, I. & Kern, S. *Nature* **537**, 89–92 (2016).
- Helm, K. P., Bindoff, N. L. & Church, J. A. *Geophys. Res. Lett.* **37**, L18701 (2010).
- Abernathy, R. P. *et al.* *Nature Geosci.* **9**, 596–601 (2016).
- Worby, A. P. *et al.* *J. Geophys. Res.* **113**, C05S92 (2008).
- Kurtz, N. T. & Markus, T. J. *Geophys. Res.* **117**, C08025 (2012).
- Holland, P. R. & Kwok, R. *Nature Geosci.* **5**, 872–875 (2012).
- Parkinson, C. L. & Cavalieri, D. J. *Cryosphere* **6**, 871–880 (2012).
- Hobbs, W. *et al.* *Glob. Planet. Change* **143**, 228–250 (2016).



50 Years Ago

Savage found that pondweeds in the presence of light stimulate spawning in *Xenopus laevis* ... This finding prompts me to report my own experience with this amphibian under more natural conditions ... At the Provincial Fisheries Institute, Lydenberg, fishponds ... are filled with water and fertilized with fowl manure in spring for the breeding of fish. Within 2 or 3 days after fertilization such ponds usually contain large numbers of *Xenopus*, which immediately start spawning, so that by the time plankton has developed the pond is teeming with larvae ... that they are attracted by fertilized water and spawn before an algal bloom develops suggests that the primary stimulus for spawning ... could be the fertilizer.
From Nature 3 September 1966

100 Years Ago

Mr. Beebe has had a wide experience of jungle-life in many lands, and hence his latest experiences in Brazil have the greater value ... Abundance of species and a relative fewness of individuals, he remarks, are pronounced characteristics of any tropical fauna ... He quickly discovered that more was to be obtained by watching particular trees ... [D]uring the space of a week of intermittent watching he obtained no fewer than seventy-six new species ... Just before leaving a brilliant idea struck Mr. Beebe ... he suddenly bethought him to fill a bag with four square feet of jungle earth, and this was examined ... while on board ship on the voyage home ... Among the captures thus made were representatives of two genera of ants new to science. There can be no doubt that important discoveries ... would accrue if this example of Mr Beebe's were generally followed in the future.

From Nature 31 August 1916

ECOLOGY

More is less

Plants compete for the same resources, such as nutrients, light and water. Because these resources are often limited, the coexistence of plant species requires the creation of trade-offs in resource use. In this issue, Harpole *et al.* report that increasing a limited nutrient in grassland can eliminate these potential trade-offs, reducing overall species diversity (W. S. Harpole *et al.* *Nature* **537**, 93–96; 2016).

The authors considered 45 grassland sites across 6 continents, and measured species diversity in response to various nutrient additions. Their results provide strong evidence for a broad ecological theory — that the availability of multiple limiting resources allows plants with different



WESTEND61/GETTY

limiting requirements to coexist.

The greater the number of limiting resources that were added, the more species were lost, although productivity and turnover improved.

The researchers argue that, by understanding the mechanisms by which diversity is lost, we might develop strategies for restoring and preserving Earth's biodiversity. [Ryan Wilkinson](#)

CANCER

Suffocation of gene expression

If a tumour outgrows its blood supply, oxygen levels in its cells decrease. It emerges that this change can alter gene expression by limiting the activity of TET enzymes, which remove methyl groups from DNA. [SEE ARTICLE P.63](#)

DAN YE & YUE XIONG

The addition of methyl groups to the DNA base cytosine leads to decreased gene expression, which has broad implications for embryonic development and tumour suppression¹. Such methylation was once considered to be irreversible, but in 2009, it was found that ten-eleven translocation (TET) enzymes could catalyse DNA demethylation². This discovery, fuelled by the finding that the gene *TET2* is frequently mutated in human blood cancers³, sparked intense interest in understanding the function and regulation of this enzyme family. Thienpont *et al.*⁴ report on page 63 that TET activity is limited by oxygen supply — revealing a general mechanism by which gene expression can be silenced in solid tumours.

TET proteins belong to a dioxygenase enzyme family, members of which depend on three cofactors for their activity: divalent iron (Fe^{2+}), the metabolite α -ketoglutarate (αKG) and oxygen⁵. Fe^{2+} in the active site of the enzyme is coordinated by αKG to split an oxygen molecule into two oxygen atoms. One oxygen atom attacks and

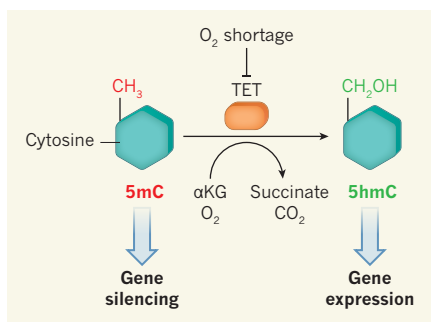


Figure 1 | Reducing TET activity through hypoxia. The addition of a methyl group (CH_3) to the DNA base cytosine to form 5-methylcytosine (5mC) can lead to silencing of many genes, including those that suppress tumour development. TET enzymes, acting with the cofactor molecules α -ketoglutarate (αKG) and oxygen, can trigger the demethylation of 5mC. In the first step of this reaction, O_2 is split into two atoms. One atom breaks a carbon–carbon bond in αKG , leading to succinate production and carbon dioxide release. The other oxidizes a carbon–hydrogen bond in CH_3 to form CH_2OH , converting 5mC to 5-hydroxymethylcytosine (5hmC), eventually leading to gene expression. Thienpont *et al.*⁴ report that a shortage of O_2 in solid tumours inhibits TET activity, leading to DNA hypermethylation.

breaks a carbon–carbon bond in αKG , leading to the conversion of the metabolite to succinate and the release of carbon dioxide. The other atom oxidizes a carbon–hydrogen bond in the enzyme's substrate (Fig. 1). In TET-mediated reactions, methylated cytosine (5-methylcytosine, 5mC) is oxidized to 5-hydroxymethylcytosine (5hmC), and further oxidation follows, eventually leading to the removal of methyl groups and so to gene expression^{6,7}.

In addition to mutations that inactivate TET genes, TET enzymes can be inactivated in tumours if their cofactors are unavailable. For example, the accumulation of αKG -competitors such as the metabolites 2-hydroxyglutarate (2-HG)^{8,9}, succinate and fumarate¹⁰ causes decreased TET activity. The discovery^{11,12} that these three metabolites accumulate in some tumours has led to the idea that cancer-promoting metabolites could have a general role in contributing to tumour development by altering the DNA-methylation landscape in cells, in much the same way that DNA damage causes cancer by altering the genomic landscape. Only a few types of cancer involve mutations in TET genes or show accumulation of αKG -competing metabolites. But the activity of TET enzymes — measured by the production of 5hmC — seems to be substantially decreased in a wide range of tumours¹³. This discrepancy has remained unexplained until now.

Solid tumours are oxygenated through blood vessels, but a tumour can rapidly outgrow its blood supply, leaving oxygen concentrations low in some regions. Thienpont *et al.* found that growing human or mouse cancer cells in such hypoxic conditions decreased 5hmC levels in some, but not all, of the cancer types they examined. Upregulation of TET gene expression

could explain the cases in which no decrease was seen.

Why do 5hmC levels decrease in most cancer-cell types in hypoxic conditions? Damaging molecules called reactive oxygen species, which could impair TET activity by reducing the amount of Fe²⁺, and metabolites that inhibit αKG such as 2-HG are known^{14,15} to be increased by hypoxia. High levels of these molecules could therefore impair TET activity. However, the authors excluded both as the cause of TET inhibition — supplements of vitamin C, which counteracts reactive oxygen species, or of αKG could not prevent 5hmC loss. Instead, analysis of enzyme kinetics predicted a 45% decrease of TET1 activity and a 52% decrease of TET2 activity in typical hypoxic tumour cells in mice. This is the first evidence that oxygen molecules are a rate-limiting factor for TET2 activity in tumours.

Cytosine methylation typically occurs at CpG dinucleotide sites, where cytosine and guanine bases are found side by side. The authors analysed CpG methylation in a few tumours. CpG sites in most genes displayed increased 5mC levels that were concomitant with reduced 5hmC levels following hypoxia, suggesting a causal link between hypoxia and DNA hypermethylation.

To test this link further, Thienpont *et al.* turned to previously established gene-expression patterns known to be a signature of hypoxia¹⁶, to assign tumours to hypoxic, normal or intermediate groups. The authors separately clustered the tumours into those that showed low, intermediate and high CpG methylation states. Hypoxic tumours predominated in the hypermethylated cluster, whereas normoxic tumours were enriched in the low-methylation cluster, providing further evidence that hypoxia leads to increased CpG methylation in tumours.

Thienpont and colleagues next found that hypoxia-linked 5hmC loss and concurrent 5mC gain were most apparent in promoter regions that drive gene expression — including the promoters of genes involved in DNA repair, the cell cycle, blood-vessel formation and cancer spread. Finally, the authors induced global loss of 5hmC *in vivo* by inducing hypoxia, and reversed this effect by deleting one copy of the oxygen-sensor gene *Phd2*, reduced function of which is known to restore tumour oxygenation¹⁷. Collectively, these results suggest that hypoxia causes TET inhibition, a reduction in 5hmC levels and DNA hypermethylation, leading to altered gene expression.

Oxygen shortage is unlikely to be the only factor that contributes to the widespread loss of 5hmC in tumours. 5hmC is a dynamic and transient modification that could be affected by changes in *TET* gene transcription, by post-translational modifications of TET proteins or even by the dynamics of DNA methylation. Thienpont and colleagues' findings also raise the question of whether hypoxia could

impair the activity of other dioxygenases that are dependent on Fe²⁺ and αKG, including those involved in DNA repair and in the demethylation of DNA-associated histone proteins.

This study also has clinical implications. Many conditions, from heart failure to stroke, can cause lasting oxygen shortage. Is TET activity impaired in these settings in ways that alter gene expression, contributing to disease progression? TET activity is frequently lost in solid tumours, but *TET* genes are rarely mutated. Could restoring TET activity in hypoxic tumours, for example by increasing levels of vitamin C, αKG or oxygen, reactivate tumour-suppressor genes that have been silenced by hypoxia-induced CpG hypermethylation?

In human tumours, drugs that inhibit blood-vessel formation have only incremental and variable benefits¹⁸, probably in part because hypoxia contributes to tumour progression and treatment resistance. In some patients, paradoxically, this treatment leads to increased tumour oxygenation and is associated with longer survival. Perhaps making an informed selection of patients on the basis of each individual's TET activity and tumour methylation status could produce therapeutic benefits. Thienpont and colleagues' study reveals a new perspective from which to further investigate the regulation of TET and other dioxygenases

that are dependent on Fe²⁺ and αKG in the development, and possibly therapeutic intervention, of hypoxia-related diseases. ■

Dan Ye is at the Institute of Biomedical Sciences and Huashan Hospital, Fudan University, Shanghai 200032, China. **Yue Xiong** is in the Department of Biochemistry and Biophysics, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, North Carolina 27599, USA.
e-mails: yedan@fudan.edu.cn;
yxiong@email.unc.edu

1. Jones, P. A. & Baylin, S. B. *Cell* **128**, 683–692 (2007).
2. Tahiliani, M. *et al.* *Science* **324**, 930–935 (2009).
3. Delhommeau, F. *et al.* *N. Engl. J. Med.* **360**, 2289–2301 (2009).
4. Thienpont, B. *et al.* *Nature* **537**, 63–68 (2016).
5. Loenarz, C. & Schofield, C. J. *Trends Biochem. Sci.* **36**, 7–18 (2011).
6. He, Y. F. *et al.* *Science* **333**, 1303–1307 (2011).
7. Ito, S. *et al.* *Science* **333**, 1300–1303 (2011).
8. Chowdhury, R. *et al.* *EMBO Rep.* **12**, 463–469 (2011).
9. Xu, W. *et al.* *Cancer Cell* **19**, 17–30 (2011).
10. Xiao, M. *et al.* *Genes Dev.* **26**, 1326–1338 (2012).
11. Selak, M. A. *et al.* *Cancer Cell* **7**, 77–85 (2005).
12. Dang, L. *et al.* *Nature* **462**, 739–744 (2009).
13. Jeschke, J., Collignon, E. & Fuks, F. *Curr. Opin. Genet. Dev.* **36**, 16–26 (2016).
14. Oldham, W. M., Clish, C. B., Yang, Y. & Loscalzo, J. *Cell Metab.* **22**, 291–303 (2015).
15. Intlekofer, A. M. *et al.* *Cell Metab.* **22**, 304–311 (2015).
16. Buffa, F. M., Harris, A. L., West, C. M. & Miller, C. J. *Br. J. Cancer* **102**, 428–435 (2010).
17. Mazzone, M. *et al.* *Cell* **136**, 839–851 (2009).
18. Jain, R. K. *Cancer Cell* **26**, 605–622 (2014).

This article was published online on 17 August 2016.

ANCIENT DNA

Muddy messages about American migration

When and by which paths did early humans migrate into America? An analysis of ancient plant and animal remains revises the timeframe during which a route may have opened between ice sheets in northwest America. [SEE ARTICLE P.45](#)

SUZANNE MCGOWAN

Towards the end of the most recent ice age, northwest America was covered by two immense ice sheets. Where the ice sheets split, there was an ice-free corridor, which was, for decades, considered to be the most probable route for the late-ice-age migration of the first humans into the Americas from Siberia. The corridor was some 1,500 kilometres long, so any path between the ice sheets would have had to develop into a viable habitat to enable humans to journey through it. On page 45, Pedersen *et al.*¹ report an analysis of ancient DNA, pollen and plant remains in lake-sediment samples from British Columbia and Alberta, Canada, at locations corresponding to stretches of the corridor. This research provides

the most complete picture yet of the timing and pattern of plant and animal development in a central 'bottleneck' region of the ice-free corridor (thought to be one of the last places in the corridor to become habitable).

It was long thought² that human colonizers from Siberia travelled across the exposed Bering land bridge and then through the ice-free corridor in western Canada, because it was the only ice-free route to the continental interior. However, subsequent evidence has led to a major re-evaluation³. It is now widely accepted that the earliest humans were present in South America by around 14,700 years ago^{4,5}, and that coalescence of the ice sheets blocked the ice-free corridor from 23,000 years ago until at least 15,000–14,000 years ago⁶. Discoveries of the earliest remains in South

America have cast doubt on whether humans could have traversed the continent in a window of, at most, 300 years. Many now favour an alternative Pacific-migration hypothesis to explain how early humans reached the Americas. This proposes that the earliest humans colonized the continent along the Pacific coastline, either by travelling along the ice-free land at coastal margins exposed by the lower sea levels, or by sea travel⁷.

Nevertheless, the corridor remains a key potential route for early migrations, particularly of the Clovis people who occupied North America from around 13,400 to 12,800 years ago, and who are associated with mammoth hunting using distinctive fluted tools^{2,3,8}. A key requirement for assessing these different route hypotheses is understanding not only when physical opening of the ice sheet occurred, but also when the flora and fauna in the corridor could have supported humans travelling northwards or southwards^{2,9}.

Pedersen *et al.* analysed deposits of up to 12,900 years old, in remnant basins of a large glacial meltwater lake that existed in one of the last corridor areas to lose ice cover. They used microscopy and radiocarbon dating to analyse pollen and plant remains from layers of lake mud, and also analysed ancient DNA¹⁰, to establish a timeline of biological changes in plants and animals. Ancient-DNA sequencing allows researchers to identify the historical presence of species such as mammals, birds and microorganisms whose remains are not visibly identifiable in sediments.

The authors chose to analyse the ancient DNA using 'shotgun' DNA sequencing, which sequences random DNA fragments in an unbiased way. This approach requires fewer assumptions to be made about which species' DNA is present than the more-specific, sequence-targeted 'barcode' method, which has been more extensively used in this field so far. As ancient-DNA-analysis techniques are starting to become more widely used¹¹ and genetic reference databases are improving, such developments are poised to revolutionize studies of ancient and complex biological remains in sediments. The ability to determine the identity of entire ancient ecological communities, including the microbes present, has great relevance for understanding their fundamental ecology, as well as providing insight into the biogeochemical processes responsible for the cycling of chemical elements through the environment.

Analysis of pollen remains by Pedersen and colleagues suggests the presence of only sparse grasses and grass-like plants known as sedge in the ice-free corridor before 12,700 years ago. However, the authors found that the key ecological successional changes occurred when the landscape was colonized by grassland vegetation known as steppe (or prairie), which included sagebrush, birch and willow. By 12,600 years ago, this steppe landscape

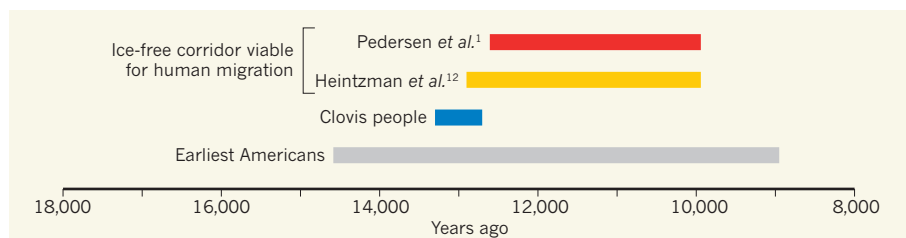


Figure 1 | Human migration into the Americas. Timeline of key events of human migrations into the Americas, including when the Clovis people were present^{2,3,8} and the timing of the earliest known human migrations into the Americas^{4,5}. One proposed migration route into the Americas from Siberia is through an ice-free corridor that opened up between two ice sheets in northwest America. The proposed ranges of dates estimated by Pedersen *et al.*¹ and Heintzman *et al.*¹² for biological viability in this ice-free corridor have implications for whether the migration of the Clovis people into the Americas could have occurred through this pathway.

supported bison; by 12,400 years ago, small mammals such as hares and voles had arrived; and mammoths, elk and bald eagles followed soon after. The presence of bison and mammoths is important because they are known to have been hunted by early Americans³, and the presence of a top predator such as the eagle indicates a productive food web. The development of coniferous forest around 10,000 years ago is also relevant, because such environments are thought to have been impassable to humans and unsuitable for the large prairie mammals that humans hunted.

The window of opportunity for human movements in the ice-free corridor is estimated by Pedersen and colleagues to have been between 12,600 and 10,000 years ago. Intriguingly, a recent publication in *Proceedings of the National Academies of Sciences* by Heintzman *et al.*¹² provides conflicting dates and suggests an earlier period during which the corridor was fully habitable (Fig. 1). Heintzman and colleagues¹² analysed DNA from organelles called mitochondria in the bones and teeth of ancient bison. The authors tested the DNA samples for the genetically distinctive signatures of bison of 'northern' and 'southern' origin. Their results indicated that the first northward movements of southern bison occurred 13,400 years ago, and that northern bison had moved south by 13,000 years ago.

However, Pedersen *et al.* suggest that the northern clade of bison could have developed before ice-sheet closure 23,000 years ago, and traversed the corridor from north to south before this time. Resolving this debate might require further consideration of whether the absence of steppe pollen and ancient DNA in the earliest sediments from the corridor region constitutes proof of absence of the species of interest, because depositional conditions in proglacial lake environments are often unstable, leading to sediment reworking and degradation¹¹.

The analysis of ancient DNA in the ice-free corridor has provided a window onto ancient worlds. Pedersen *et al.* focus on the role of the food web in the steppe environment in supporting early humans, but, as they acknowledge, the ancient-DNA records are

incomplete. For example, archaeological records from the nearby Charlie Lake Cave indicate that fish and waterfowl were key dietary components of dwellers close to the lake from 12,700 years ago¹³, but the authors found no genetic evidence for these species at this time. By contrast, the ancient DNA fills in gaps in the pollen record, indicating that after 12,400 years ago, poplar and willow trees were more common than previously estimated⁹.

Together, the availability of wood and freshwater resources might have implications for travel in this region — the vast glacial lakes that covered parts of the corridor might be considered more of an opportunity for water transport than an impediment to passage. The Pacific-migration hypothesis implies earlier use of watercraft than has previously been assumed⁷. Further investigation of the hypotheses for human migration into the Americas will require close integration of studies analysing archaeology, genetics and ancient environments, which should, in turn, identify pathways for developing more-complete interpretations of sedimentary ancient DNA¹⁴. ■

Suzanne McGowan is at the School of Geography, University of Nottingham, Nottingham NG7 2RD, UK.
e-mail: suzanne.mcgowan@nottingham.ac.uk

- Pedersen, M. W. *et al.* *Nature* **537**, 45–49 (2016).
- Ives, J. W., Froese, D., Supernant, K. & Yanicki, G. in *Paleoamerican Odyssey* (eds Graf, K. E., Ketron, C. V. & Waters, M. R.) 149–170 (2014).
- Goebel, T., Waters, M. R. & O'Rourke, D. H. *Science* **319**, 1497–1502 (2008).
- Dillehay, T. D. *et al.* *Science* **320**, 784–786 (2008).
- Gilbert, M. T. P. *et al.* *Science* **320**, 786–789 (2008).
- Dyke, A. S. *Dev. Quat. Sci.* **2**, 373–424 (2004).
- Erlandson, J. M. in *Paleoamerican Odyssey* (eds Graf, K. E., Ketron, C. V. & Waters, M. R.) 127–132 (2014).
- Waters, M. R., Stafford, T. W., Kooyman, B. & Hills, L. V. *Proc. Natl Acad. Sci. USA* **112**, 4263–4267 (2015).
- Mandryk, C. A. S., Josenhans, H., Fedje, D. W. & Mathewes, R. W. *Quat. Sci. Rev.* **20**, 301–314 (2001).
- Pedersen, M. W. *et al.* *Phil. Trans. R. Soc. B* **370**, 20130383 (2015).
- Birks, H. J. B. & Birks, H. H. *New Phytol.* **209**, 499–506 (2016).
- Heintzman, P. D. *et al.* *Proc. Natl Acad. Sci. USA* **113**, 8057–8063 (2016).
- Driver, J. C. *Can. J. Earth Sci.* **25**, 1545–1553 (1988).
- Seddon, A. W. R. *et al.* *J. Ecol.* **102**, 256–267 (2014).

This article was published online on 10 August 2016.

Postglacial viability and colonization in North America's ice-free corridor

Mikkel W. Pedersen¹, Anthony Ruter¹, Charles Schweger², Harvey Friebe², Richard A. Staff³, Kristian K. Kjeldsen^{1,4}, Marie L. Z. Mendoza¹, Alwynne B. Beaudoin⁵, Cynthia Zutter⁶, Nicolaj K. Larsen^{1,7}, Ben A. Potter⁸, Rasmus Nielsen^{1,9,10}, Rebecca A. Rainville¹¹, Ludovic Orlando¹, David J. Meltzer^{1,12}, Kurt H. Kjær¹ & Eske Willerslev^{1,13,14}

During the Last Glacial Maximum, continental ice sheets isolated Beringia (northeast Siberia and northwest North America) from unglaciated North America. By around 15 to 14 thousand calibrated radiocarbon years before present (cal. kyr BP), glacial retreat opened an approximately 1,500-km-long corridor between the ice sheets. It remains unclear when plants and animals colonized this corridor and it became biologically viable for human migration. We obtained radiocarbon dates, pollen, macrofossils and metagenomic DNA from lake sediment cores in a bottleneck portion of the corridor. We find evidence of steppe vegetation, bison and mammoth by approximately 12.6 cal. kyr BP, followed by open forest, with evidence of moose and elk at about 11.5 cal. kyr BP, and boreal forest approximately 10 cal. kyr BP. Our findings reveal that the first Americans, whether Clovis or earlier groups in unglaciated North America before 12.6 cal. kyr BP, are unlikely to have travelled by this route into the Americas. However, later groups may have used this north-south passageway.

Understanding the postglacial emergence of an unglaciated and biologically viable corridor between the retreating Cordilleran and Laurentide ice sheets is a key part of the debate on human colonization of the Americas^{1–3}. The opening of the ice-free corridor, long considered the sole entry route for the first Americans, closely precedes the ‘abrupt appearance’ of Clovis, the earliest widespread archaeological complex south of the ice sheets at ~13.4 cal. kyr BP^{4,5}. This view has been challenged by recent archaeological evidence that suggests people were in the Americas by at least 14.7 cal. kyr BP^{6,7}, and possibly several millennia earlier⁸. Whether this earlier presence relates to Clovis groups remains debated⁹. Regardless, as it predates all but the oldest estimates for the opening of the ice-free corridor^{10,11}, archaeological attention has shifted to the Pacific coast as an alternative early entry route into the Americas^{1,11}. Yet, the possibility of a later entry in Clovis times through an interior ice-free corridor remains open^{1,9,12}.

Whether the ice-free corridor could have been used for a Clovis-age migration depends on when it became biologically viable. However, determining this has proven difficult because radiocarbon and luminescence dating of ice retreat yield conflicting estimates for when the corridor opened, precluding precise reconstruction of deglaciation chronology^{10,13–17}. Once the landscape was free of ice and meltwater, it was open for occupation by plants and animals, including those necessary for human subsistence. On the basis of studies on modern glaciers¹⁸, the onset of biological viability could have been brief (for example, a few decades) if propagules were available in adjacent areas, and assuming they were capable of colonizing what would have been a base-rich (high pH) and nitrogen-poor, soil substrate (such as nitrogen-fixing plants like *Shepherdia canadensis* (buffaloberry)).

Establishment of biota within the corridor region must have varied locally depending on the rate and geometry of ice retreat, the extent of landscape flooding under meltwater lakes, and the proximity of plant

and animal taxa and their dispersal mechanisms^{1,19,20}. Some areas were habitable long before others. Although the corridor's deglaciation history was complex, broadly speaking it first opened from its southern and northern ends, leaving a central bottleneck that extended from approximately 55°N to 60°N^{1,10,13–15,21}. On the basis of currently available geological evidence, this was the last segment to become ice free and re-colonized by plants and animals^{1,13,22–24}.

Although palynological and palaeontological data can be used to help study the opening of the corridor region, these are limited in several respects. First, not all vegetation, particularly pioneering forbs and shrubs, produce pollen and macrofossils with good preservation potential that will be detectable in available depositional locales. Hence, timing of plants' appearance may be underestimated. Second, pollen can disperse over long distances and have limited taxonomic resolution, differential preservation, and variable production rates, all of which can bias vegetation reconstruction²⁵. Third, fossil evidence for initial large mammal populations that dispersed into the newly opened corridor is sparse. The fossil remains suggest the presence of bison, horse and mammoth, and probably some camel, muskox and caribou^{26,27}. Yet, the oldest vertebrate remains after the Last Glacial Maximum are no older than ~13.5 cal. kyr BP², and those specimens are found outside the bottleneck region^{1,3,26,28,29}. These animals would have been the source populations to recolonize the newly opened landscape, and thus their presence within the bottleneck region can indicate when the corridor became a viable passageway over its entirety.

Samples and analytical approaches

To overcome current limitations of the palaeoecological record, and develop a more precise chronology for the opening and biological viability of corridor's bottleneck region, we collected nine lake sediment cores from Charlie Lake and Spring Lake in the Peace River drainage

¹Centre for GeoGenetics, Natural History Museum, University of Copenhagen, Copenhagen 1350, Denmark. ²Department of Anthropology, University of Alberta, Edmonton, Alberta T6G 2H4, Canada. ³School of Archaeology, University of Oxford, Oxford OX1 3QY, UK. ⁴Department of Earth Sciences, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada. ⁵Royal Alberta Museum, Edmonton, Alberta T5N 0M6, Canada. ⁶Department of Anthropology, MacEwan University, Edmonton, Alberta T5J 4S2, Canada. ⁷Department of Geoscience, Aarhus University, 8000 Aarhus, Denmark. ⁸Department of Anthropology, University of Alaska Fairbanks, Fairbanks, Alaska 99775, USA. ⁹Department of Integrative Biology, University of California, Berkeley, California 94720-3140, USA. ¹⁰Department of Biology, University of Copenhagen, 2200 Copenhagen, Denmark. ¹¹Department of Archaeology, University of Calgary, Calgary, Alberta T2N 1N4, Canada. ¹²Department of Anthropology, Southern Methodist University, Dallas, Texas 75275, USA. ¹³Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK. ¹⁴Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK.

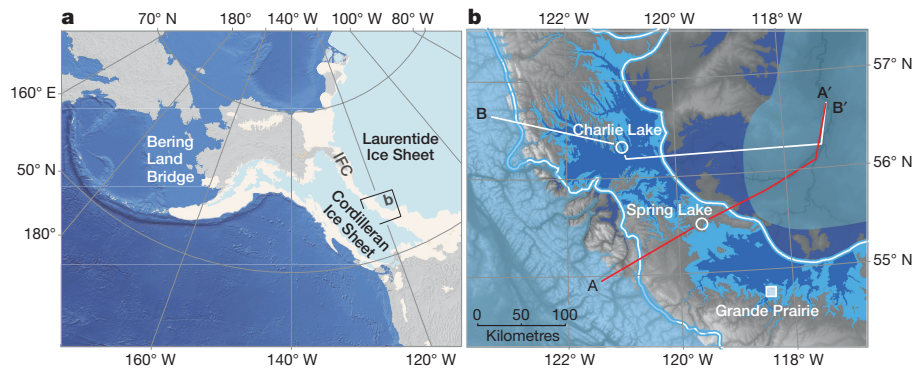


Figure 1 | Setting and study area. During the Last Glacial Maximum, the Laurentide Ice Sheet and the Cordilleran Ice Sheet coalesced in western mid-Canada creating a physical barrier to north–south migration. Following the Last Glacial Maximum, the ice retreated creating an ice-free corridor (IFC). **a**, Ice extent¹⁰ during two periods, Last Glacial Maximum 21.4 cal. kyr BP (off-white) and Late Pleistocene 14.1 cal. kyr BP (light-blue).

basin (Fig. 1). These are remnants of Glacial Lake Peace, which formed as the Laurentide Ice Sheet began to retreat in this region around 15 to 13.5 cal. kyr BP and blocked eastward draining rivers^{10,13–15,21} (Extended Data Fig. 1). Glacial Lake Peace flooded the gap between the ice fronts until about 13 cal. kyr BP, sometime after which Charlie and Spring lakes became isolated¹³. Thus, this area was amongst the last segments of the corridor to open and is pivotal to understanding its history as a biogeographic passageway^{1,13,14,16,22,24}.

Of the nine cores obtained from Charlie Lake and Spring Lake, one from each lake predates the Pleistocene to Holocene transition, the oldest dating to ~12.9 cal. kyr BP (modelled age). We sampled the cores from both lakes for magnetic susceptibility, pollen^{30,31}, micro- and macrofossils, including ¹⁴C-dateable material for subsequent robust Bayesian age-depth modelling (Fig. 2, Methods, Extended Data Figs 2–4 and Supplementary Information). In addition, we obtained environmental DNA (eDNA)³², representing molecular fossils of local organisms derived from somatic tissues, urine and faeces³³, but rarely pollen³⁴. eDNA complements traditional pollen and macrofossil studies³⁵, and is especially useful for establishing the likelihood that a taxon occurred within a particular time period^{36,37}. Furthermore, eDNA enables identification of taxa even in the absence of micro- and macrofossil material, thus improving the resolution of taxonomic richness surveys³⁶. However, amplification of short and taxonomically informative DNA metabarcodes³⁸ can be biased towards taxa targeting³⁵. We used shotgun sequencing of the full metagenome in the DNA extracts to reveal the whole diversity of taxonomic groups present in the sediment³⁹ (Fig. 2, Methods, Extended Data Figs 5 and 6 and Supplementary Information). We confirmed the sequences identified as ancient by quantifying DNA damage⁴⁰, and found the DNA damage levels to accumulate with age (Pearson correlation coefficient = 0.663, *P* value = 0.00012) (Methods and Extended Data Fig. 7a, b).

Biological succession within the corridor bottleneck

The basal deposit in the Charlie Lake core is proglacial gravel, previously reported from the area²², above which are laminated lacustrine sediments, principally composed of silt-sized grains²⁴ (Extended Data Fig. 2). We interpret these as deposits from Glacial Lake Peace Stage IV (ref. 13), the >15,000 km² proglacial lake that covered the Peace River area of northeastern British Columbia and northwestern Alberta. A subsequent lithological change from silt to sandy organic rich mud (gyttja) at the onset of Holocene, around 11.6 cal. kyr BP, reflects a change in sediment source and lake productivity we interpret as Charlie Lake becoming isolated from Glacial Lake Peace (Fig. 1). This is followed by a decrease in pollen influx in both lake records at ~11.5 cal. kyr BP that coincides with an increase in pre-Quaternary palynomorphs. At Charlie Lake there is then a marked increase in

b, Topography of the Peace River basin with Glacial Lake Peace Phase III (white lines with blue outlines) and Phase IV¹³ (light-blue and dark-blue) at around 14.1 cal. kyr BP and 13 cal. kyr BP, respectively. The red and white lines mark topographic transects of the lakes which in relation to the four phases of Glacial Lake Peace¹³ is found in Extended Data Fig. 1.

pollen influx at ~11.3 cal. kyr BP. We interpret these fluctuations as responses of a highly dynamic landscape to paraglacial and aeolian redepositional processes.

Our palynological and eDNA-based taxonomic identifications, respectively, reveal the development of biota in the regional and local environment surrounding each lake (Fig. 2, Extended Data Figs 3–6). Prior to ~12.6 cal. kyr BP (Charlie Lake, pollen zone I, ~13 to 12.6 cal. kyr BP), the bottleneck area appears to have been largely unvegetated, receiving low pollen influx (<50 grains cm⁻² y⁻¹) with little organic content (incoherent/coherent ratio) and low DNA concentrations (<5 ng per g of sediment). During the later phases of Glacial Lake Peace, both pollen and eDNA indicate grasses and sedges were early colonizers. Charlie Lake pollen zone II (~12.6 to 11.6 cal. kyr BP) contains evidence of steppe vegetation, including *Artemisia* (sagebrush), Asteraceae (sunflower family), Ranunculaceae (buttercup family), Rosaceae (rose family, rosids in eDNA), *Betula* (birch), and *Salix* (willow). A similar plant community is recorded at Spring Lake (pollen zone 1), with substantial abundances of *Populus* and *S. canadensis*, probably due to elevation differences and because by this time Spring Lake was no longer part of the Glacial Lake Peace system.

eDNA indicates the steppe vegetation supported a variety of animals including *Bison* which appear at ~12.5 cal. kyr BP, and *Microtus* (vole) and *Lepus* (jackrabbit) by ~12.4 cal. kyr BP (Fig. 3). After 12.4 cal. kyr BP, *Populus* trees became more dominant and *Cervus* (elk), *Haliaeetus* (bald eagle) and *Alces* (moose) appear in the eDNA record. The productivity of the bottleneck increased to a peak at ~11.6 cal. kyr BP. The presence of *Esox* (pike), a top aquatic predator, implies that by ~11.7 cal. kyr BP, a fish community was already established. After 11.6 cal. kyr BP, *Picea* (spruce), *Pinus* (pine) and *Betula* pollen increased in the Charlie Lake pollen record, reflecting the establishment of boreal forest.

Around 11.5 cal. kyr BP, a distinct decline occurred in pollen influx at both lakes. High abundance of *Botryococcus* (green algae) in each is probably a response to changing nutrient sources, lake chemistry, sediment input and possibly reduced turbidity following isolation of these basins from Glacial Lake Peace⁴¹. *Botryococcus* dominated the early Holocene sequence in Spring Lake (11.7–11.5 cal. kyr BP) but declined relative to *Pediastrum* (green algae) after 11.0 cal. kyr BP, consistent with eutrophication in a more productive ecosystem. Pollen and plant macrofossils indicate *Alnus* (alder) was in the vicinity of Spring Lake at about 7.0 cal. kyr BP, although it is not evident in eDNA until approximately 5.5 cal. kyr BP.

We used non-metric multi-dimensional scaling (NMDS) based on Bray–Curtis similarity measures to explore whether the eDNA plant communities, excluding algae, reflect the pollen data (Fig. 2b, d).

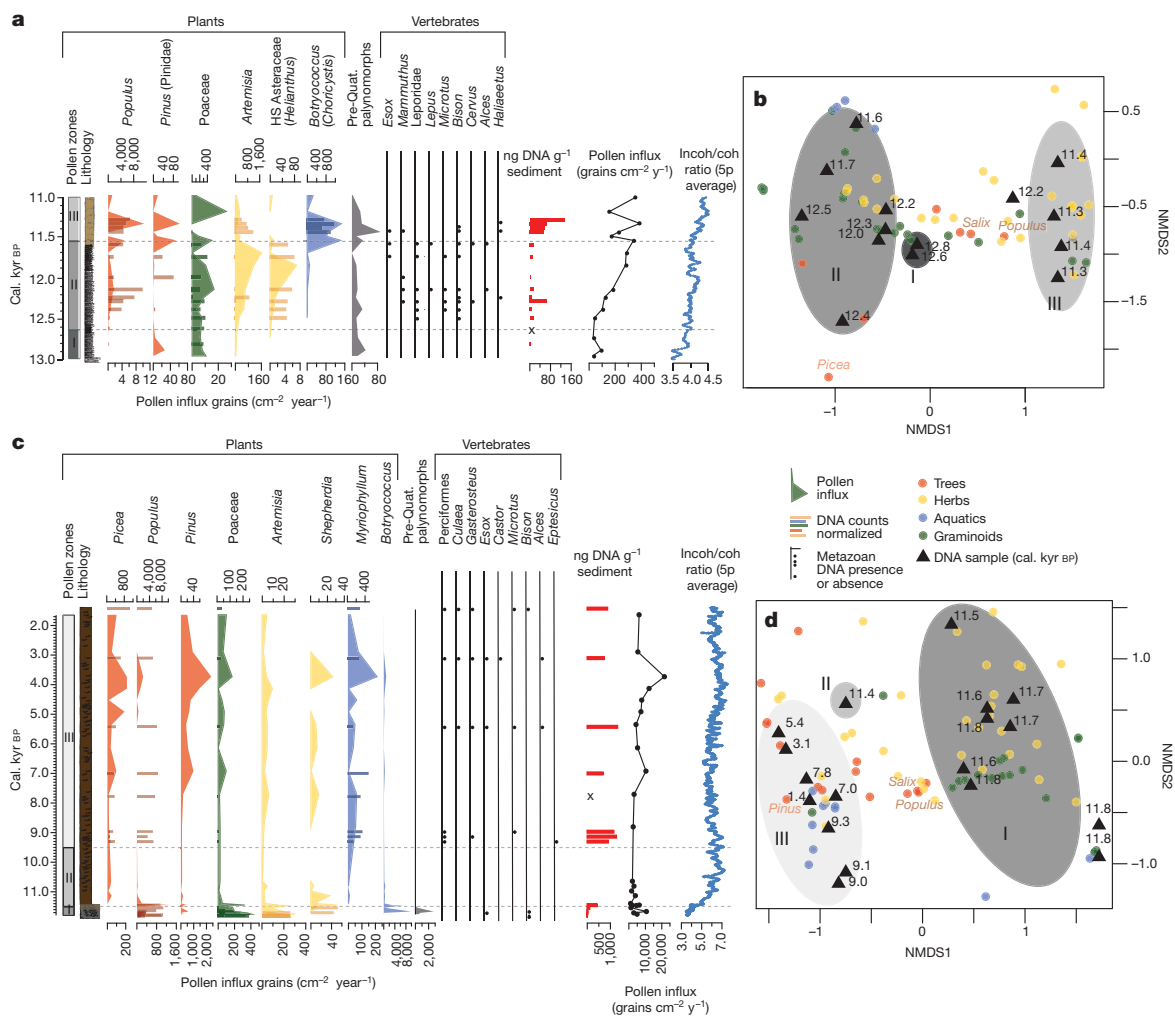


Figure 2 | Selected pollen, DNA and biometrical results. a, c, Pollen are presented as influx (area) and DNA taxa presented with normalized counts (bars). HS asteraceae, high spike asteraceae. Metazoans are presented with bullet points indicating their presence. The 5 point average (5p) of the incoherent/coherent (incoh/coh) ratio is derived from the X-ray fluorescence results and an increasing ratio represents increased organic

content. **b, d**, Non-metric multi-dimensional scaling plots; grey ellipses marked I, II, and III encircle the samples corresponding to the respective CONIIC pollen zonation. Coloured dots indicate each taxon identified. The coloured categories are identical to the pollen and DNA taxa in Charlie Lake (**a**), and Spring Lake (**c**).

In eDNA samples, the first NMDS axis matches the clear separation between major pollen zones at Spring Lake and Charlie Lake. The only exception is represented by the 12.2 cal. kyr BP sample at Charlie Lake, which does not cluster with other samples of similar age (~12.6–11.6 cal. kyr BP) but is closer to the arboreal and younger samples from pollen zone I. Nevertheless, consistency between the main pollen zones and clustering of eDNA samples confirms that large ecological changes found in pollen records can be identified using eDNA.

Despite good conformity between palynological and eDNA data, some discrepancies suggest these proxies are variably affected by a plant's reproductive process and taphonomic history (see Supplementary Information). The most notable of these discrepancies is the *Populus* record. In Charlie Lake, its pollen and eDNA signals are congruent from ~ 11.6 – 11.2 cal. kyr BP, whereas earlier (~ 12.4 – 12.1 cal. kyr BP) the eDNA signal for *Populus* is more pronounced. In Spring Lake, *Populus* pollen only occurs towards the base of the record and in upper zone III, whereas in the eDNA record it is abundant throughout. This discrepancy is probably due to *Populus* reproducing vegetatively, and its notoriously low detection rates and poor pollen preservation, which often render it palynologically 'silent'⁴². The eDNA reveals that poplar was probably more abundant in the regional vegetation than has previously been shown with palynology. This has important implications for human occupation as poplar would

have provided wood for fuel, shelter, and tools, as well as browse feeding for animals.

The differences between the pollen and eDNA evidence for plants might also reflect dispersal factors. Wind-dispersed pollen is more likely to be encountered in lake-based pollen records, whereas predominantly insect-pollinated taxa are less likely to settle in lake sediments and be detected. Many willows (*Salix* spp.), for example, are insect pollinated. Their pollen is present in low percentage (5%) in zone II in Charlie Lake, but in higher abundance in zones II and III in the eDNA record (Extended Data Figs 5 and 6). This suggests the eDNA comes more from macrofossils and plant debris than from pollen.

The eDNA record also detects taxa not present in fossil bone assemblages, including terrestrial and aquatic vertebrates. In particular, it identifies top-level aquatic (*Esox*) and avian (*Haliaeetus*) predators, which indicate a rich supporting community at lower trophic levels. *Cervus* is evident in the Charlie Lake record at about 11.5 cal. kyr BP, whereas its earliest fossil remains from the area date to about 10.2 cal. kyr BP⁴³. Small mammals, such as *Microtus* are documented in the Charlie Lake eDNA at 12.4 cal. kyr BP confirming the *Microtus* colony found just west of Charlie Lake, at Bear Flats⁴³. Yet, there are also notable absences in eDNA compared to the vertebrate record. For example, faunal remains from the adjacent Charlie Lake Cave, dated to ~12.4 cal. kyr BP⁴⁴ are rich in waterfowl and other birds and fish not

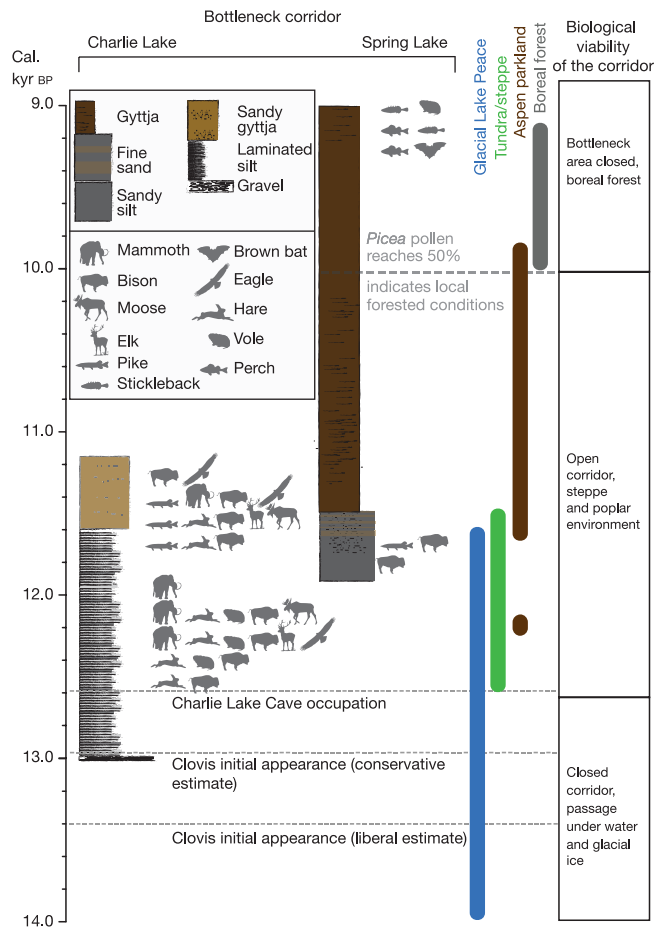


Figure 3 | Ecological interpretation and implications of this study.

Timeline of the biology in the bottleneck area linking it with evidence of human occupation and the first appearance of Clovis technology (see also Fig. 4). Grey animal silhouettes are vertebrate genera that were identified by environmental DNA in both lake cores.

detected by eDNA. In the Spring Lake eDNA record, *Castor* (beaver) appears between 5.4 and 3 cal. kyr BP, whereas evidence from Wood Bog⁴⁵ ~60 km to the south suggests that the beaver was part of the local fauna since at least 11 cal. kyr BP.

When the evidence from these multiple proxies is combined, it provides a more robust record of the presence of plants and animals than any single indicator. It is, of course, possible that some taxa arrived on the landscape earlier and escaped detection, thus appearing absent. However, there was only a narrow window of time between when the bottleneck region was beneath the waters of Glacial Lake Peace and

impassable, and when these proxies first detect the presence of plants and animals. The eDNA data are particularly important for indicating the earliest occurrence of terrestrial fauna in the bottleneck region, particularly the game animals that would have been key subsistence resources for hunter-gatherers⁴⁶.

Discussion

Although ice sheet retreat led to the corridor physically opening in the bottleneck region starting around 15–14 cal. kyr BP¹⁰, deglaciation was followed by regional inundation below the waters of Glacial Lake Peace for perhaps up to 2,000 years¹³. By around 12.6 cal. kyr BP the ice sheets were several hundred kilometres apart and the landscape had become vegetated. Large and small animals came in soon thereafter, around 12.5 cal. kyr BP, making the corridor capable of supplying the biotic resources, including high-ranked prey such as bison, required by human foragers for the 1,500 km traverse⁴⁷. This result is consistent with the recent finding that the oldest of the southern bison clade specimens (clades 1a and 2b) found north of the bottleneck region postdates 12.5 cal. kyr BP, though not with the finding that it opened earlier³ (see Supplementary Information).

From our findings, it follows that an ice-free corridor was unavailable to those groups who appear to have arrived in the Americas south of the continental ice sheets by 14.7 cal. kyr BP^{6,7}, and also opened too late to have served as an entry route for the ancestors of Clovis who were present by 13.4 cal. kyr BP^{1,9}. Not surprisingly, the earliest archaeological presence in the Peace River region, at Charlie Lake Cave (Fig. 3) and Saskatoon Mountain^{45,47}, postdates 12.6 cal. kyr BP. More striking, once opened, the corridor was not used just for southbound movement: archaeological evidence suggests that people were moving north as well, potentially renewing contact between groups that had been separated for millennia^{1,9}. Bison³ were also colonizing the corridor and moving north and south; it is uncertain whether other species, such as elk² and brown bears⁴⁸, were moving similarly.

More broadly, although Clovis people may yet be shown to represent an independent migration separate from the peoples present here by 14,700 cal. kyr BP, they must have descended from a population that entered the Americas via a different route than the ice-free corridor. This conclusion is relevant to the recent finding⁴⁹ that ancestral Native Americans diverged into southern and northern branches ~13 cal. kyr BP (95% confidence interval of 14.5–11.5 cal. kyr BP). This implies that if that split occurred north of the ice sheets, there must have been two pulses of migration to the south. As the Anzick infant's genome, dated to 12.6 cal. kyr BP and associated with Clovis artefacts, is part of the southern branch⁵⁰, its ancestors must have travelled via the coast. However, this does not preclude the possibility that ancestors of the northern branch left Alaska later, through a then-viable ice-free corridor. Alternatively, if the divergence occurred in unglaciated North America, as recently proposed⁴⁹, it implies a single ancestral population came via the coast. It further raises the possibility that the

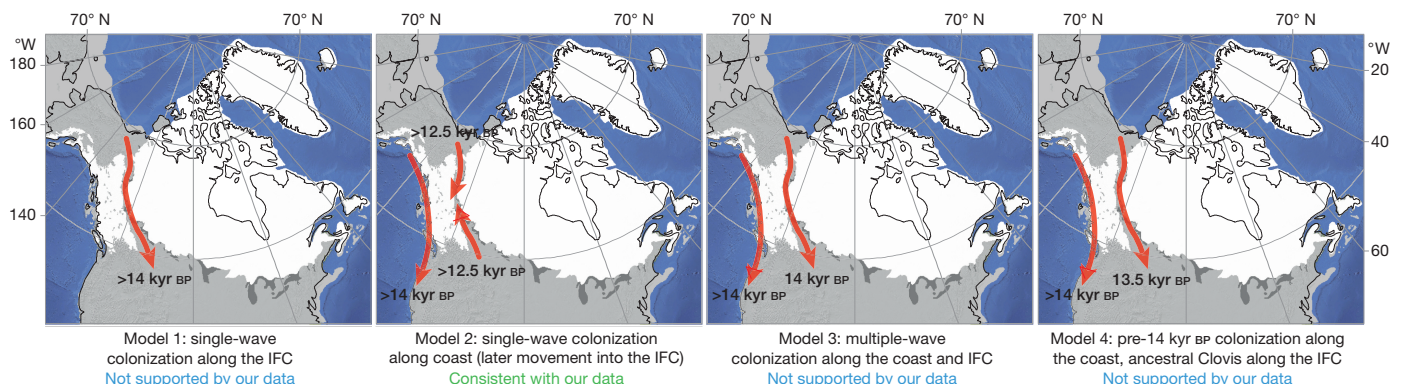


Figure 4 | Colonization models. Comparison of models of Paleoindian colonization (number of pulses, timing, and route(s)) that are supported or rejected by our data. All ages are in calibrated years before present.

northern branch—the descendants occupying Alaska today—made their way north to Alaska via the corridor after 12.6 cal. kyr BP. Further investigations of ancient DNA may help resolve this issue.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 February; accepted 7 July 2016.

Published online 10 August 2016.

- Ives, J. W., Froese, D., Supernant, K. & Yanicki, G. M. *Paleoamerican Odyssey* 149–169 (Texas A & M Univ. Press, 2014).
- Meiri, M. *et al.* Faunal record identifies Bering isthmus conditions as constraint to end-Pleistocene migration to the New World. *Proc. R. Soc. Lond. B* **281**, 20132167 (2013).
- Heintzman, P. D. *et al.* Bison phylogeography constrains dispersal and viability of the Ice Free Corridor in western Canada. *Proc. Natl Acad. Sci. USA* **113**, 8057–8063 (2016).
- Ferring, C. R. *The Archaeology and Paleoecology of the Aubrey Clovis Site (41DN479)* (Denton County, 2001).
- Sanchez, G. *et al.* Human (Clovis)-gomphothere (*Cuvieronius* sp.) association ~ 13,390 calibrated yBP in Sonora, Mexico. *Proc. Natl Acad. Sci. USA* **111**, 10972–10977 (2014).
- Dillehay, T. D. *et al.* Monte Verde: seaweed, food, medicine, and the peopling of South America. *Science* **320**, 784–786 (2008).
- Gilbert, M. T. P. *et al.* DNA from pre-Clovis human coprolites in Oregon, North America. *Science* **320**, 786–789 (2008).
- Dillehay, T. D. *et al.* New archaeological evidence for an early human presence at Monte Verde, Chile. *PLoS One* **10**, e0141923 (2015).
- Meltzer, D. J. *First Peoples in a New World: Colonizing Ice Age America* (Univ. California Press, 2009).
- Dyke, A. S. An outline of North American deglaciation with emphasis on central and northern Canada. *Quaternary Glaciations: Extent and Chronology* 371–406 (Elsevier, 2004).
- Dixon, E. J. Late Pleistocene colonization of North America from northeast Asia: new insights from large-scale paleogeographic reconstructions. *Quat. Int.* **285**, 57–67 (2013).
- Madsen, D. B. A framework for the initial occupation of the Americas. *PaleoAmerica* **1**, 217–250 (2015).
- Hickin, A. S., Lian, O. B., Levson, V. M. & Cui, Y. Pattern and chronology of glacial Lake Peace shorelines and implications for isostasy and ice-sheet configuration in northeastern British Columbia, Canada. *Boreas* **44**, 288–304 (2015).
- Hickin, A. S., Lian, O. B. & Levson, V. M. Coalescence of late Wisconsinan Cordilleran and Laurentide ice sheets east of the Rocky Mountain foothills in the Dawson Creek region, northeast British Columbia, Canada. *Quat. Res.* **85**, 409–429 (2016).
- Munyikwa, K., Feathers, J. K., Rittenour, T. M. & Shrimpton, H. K. Constraining the Late Wisconsinan retreat of the Laurentide ice sheet from western Canada using luminescence ages from postglacial aeolian dunes. *Quat. Geochronol.* **6**, 407–422 (2011).
- White, J. M., Mathewes, R. W. & Mathews, W. H. Late Pleistocene chronology and environment of the ‘ice-free corridor’ of northwestern Alberta. *Quat. Res.* **24**, 173–186 (1985).
- James Dixon, E. Human colonization of the Americas: timing, technology and process. *Quat. Sci. Rev.* **20**, 277–299 (2001).
- Viereck, L. A. Plant succession and soil development on gravel outwash of the Muldrow Glacier, Alaska. *Ecol. Monogr.* **36**, 181–199 (1966).
- Mandryk, C. A. S., Josenhans, H., Fedje, D. W. & Mathewes, R. W. Late Quaternary paleoenvironments of northwestern North America: implications for inland versus coastal migration routes. *Quat. Sci. Rev.* **20**, 301–314 (2001).
- Stokes, C. R., Margold, M., Clark, C. D. & Tarasov, L. Ice stream activity scaled to ice sheet volume during Laurentide Ice Sheet deglaciation. *Nature* **530**, 322–326 (2016).
- Gowan, E. J. An assessment of the minimum timing of ice free conditions of the western Laurentide Ice Sheet. *Quat. Sci. Rev.* **75**, 100–113 (2013).
- Mathews, W. H. Quaternary stratigraphy and geomorphology of Charlie Lake (94a) map area, British Columbia. *Canadian Geolocal Survey* <http://dx.doi.org/10.4095/104544> (1978).
- White, J. M., Mathewes, R. W. & Mathews, W. H. Radiocarbon dates from Boone Lake and their relation to the ‘ice-free corridor’ in the Peace River district of Alberta, Canada. *Can. J. Earth Sci.* **16**, 1870–1874 (1979).
- Hartman, G. & Clague, J. J. Quaternary stratigraphy and glacial history of the Peace River valley, northeast British Columbia. *Can. J. Earth Sci.* **45**, 549–564 (2008).
- Birks, H. H. & Birks, H. J. B. *Quaternary Palaeoecology* (Edward Arnold, 1980).
- Jass, C. N., Burns, J. A. & Milot, P. J. Description of fossil muskoxen and relative abundance of Pleistocene megafauna in central Alberta. *Can. J. Earth Sci.* **48**, 793–800 (2011).
- Kooyman, B., Hills, L. V., McNeil, P. & Tolman, S. Late Pleistocene horse hunting at the Wally’s Beach site (DhPg-8), Canada. *Am. Antiq.* **71**, 101–121 (2006).
- Burns, J. A. Mammalian faunal dynamics in Late Pleistocene Alberta, Canada. *Quat. Int.* **217**, 37–42 (2010).
- Kooyman, B., Hills, L., Tolman, S. & McNeil, P. Late Pleistocene western camel (*Camelops hesternus*) hunting in southwestern Canada. *Am. Antiquity* **77**, 115–124 (2012).
- Faegri, K., Kaland, P. E. & Krzywinski, K. *Textbook of Pollen Analysis* 1–328 (John Wiley and Sons, 1990).
- Bennett, K. D. *Documentation for PSIMPOLL 4.10 and PSCOMB 1.03*. 1–127 (Univ. of Uppsala, Sweden, 2005).
- Willerslev, E. *et al.* Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science* **300**, 791–795 (2003).
- Pedersen, M. W. *et al.* Ancient and modern environmental DNA. *Proc. R. Soc. Lond. B* <http://dx.doi.org/10.1098/rstb.2013.0383> (2015).
- Pedersen, M. W. *et al.* A comparative study of ancient environmental DNA to pollen and macrofossils from lake sediments reveals taxonomic overlap and additional plant taxa. *Quat. Sci. Rev.* **75**, 161–168 (2013).
- Parducci, L. *et al.* Molecular- and pollen-based vegetation analysis in lake sediments from central Scandinavia. *Mol. Ecol.* **22**, 3511–3524 (2013).
- Haile, J. *et al.* Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proc. Natl Acad. Sci. USA* **106**, 22352–22357 (2009).
- Parducci, L. *et al.* Glacial survival of boreal trees in northern Scandinavia. *Science* **335**, 1083–1086 (2012).
- Valentini, A., Pompanon, F. & Taberlet, P. DNA barcoding for ecologists. *Trends Ecol. Evol.* **24**, 110–117 (2009).
- Coissac, E., Hollingsworth, P. M., Lavergne, S. & Taberlet, P. From barcodes to genomes: extending the concept of DNA barcoding. *Mol. Ecol.* **25**, 1423–1428 (2016).
- Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684 (2013).
- Hickman, M. & White, J. Late Quaternary palaeoenvironment of Spring Lake, Alberta, Canada. *J. Paleolimnol.* **2**, 305–317 (1989).
- Godwin, H. Pollen analysis, an outline of the problems and potentialities of the method. Part I. Technique and interpretation. *New Phytol.* **33**, 278–305 (1934).
- Hebda, R. J., Burns, J. A., Geertsema, M. & Timothy Jull, A. J. AMS-dated late Pleistocene taiga vole (*Microtus xanthognathus*) from northeast British Columbia, Canada: a cautionary lesson in chronology. *Can. J. Earth Sci.* **45**, 611–618 (2008).
- Harrington, C. R. Quaternary cave faunas of Canada: a review of the vertebrate remains. *J. Caves Karst Stud.* **73**, 162–180 (2009).
- Beaudoin, A. B., Wright, M. & Ronaghan, B. Late quaternary landscape history and archaeology in the ‘Ice-Free Corridor’: Some recent results from Alberta. *Quat. Int.* **32**, 113–126 (1996).
- Potter, B. A., Holmes, C. & Yesner, D. R. *Paleoamerican Odyssey* 81–103 (Texas A & M Univ. Press, 2014).
- Driver, J. C. & Vallières, C. The Palaeoindian bison assemblage from Charlie Lake Cave, British Columbia. *Can. J. Archaeol.* **32**, 239–257 (2008).
- Mathews, P., Burns, J., Weinstock, J. & Hofreiter, M. Pleistocene brown bears in the mid-continent of North America. *Science* **306**, 1150 (2004).
- Raghavan, M. *et al.* Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, <http://dx.doi.org/10.1126/science.aab3884> (2015).
- Rasmussen, M. *et al.* The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**, 225–229 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. L. Tobiaz, T. Murchie, G. Carroll, S. Overballe-Petersen, M. Reasoner, K. Walde, D. Wilson, F. Malekani, A. Freeman, J. Holm, St John’s College in Cambridge, and the Danish National Sequencing Centre for help and support. The Fig. 1b map contains a digital elevation model licensed under the Open Government Licence – Canada (<http://open.canada.ca/en/open-government-licence-canada>). This study was supported by the Danish National Research Foundation (DNRF94), the Lundbeck Foundation and KU2016.

Author Contributions E.W. initiated and led the study. M.W.P., K.H.K., and E.W. designed and conducted the study. A.R., C.S., C.Z. and H.F. processed and counted pollen and macrofossils. R.A.S. performed the ^{14}C dating and Bayesian age modelling. N.K.L. and R.A.R. scanned cores for X-ray fluorescence and magnetic susceptibility. K.K.K., M.W.P. and K.H.K. performed the cartographic analysis and representation. M.L.Z.M. and M.W.P. processed and analysed the metabarcoding data set. M.W.P. performed the molecular work under supervision by L.O. and E.W. M.W.P., C.S., A.B.B., B.A.P., D.J.M., K.H.K. and E.W. did the main interpretations of the results, with additional statistical analysis from R.N. M.W.P., D.J.M. and E.W. wrote the paper with input from all authors.

Author Information DNA sequence data are available through the European Nucleotide Archive under accession number PRJEB14494 and bioinformatics scripts are available at (<https://github.com/ancient-eDNA/Holi>). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.W. (ewillerslev@snm.ku.dk).

Reviewer Information

Nature thanks P. Gibbard, S. McGowan, A. P. Roberts and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Sediment sampling. We obtained 23 sediment cores from 8 different lakes by using a percussion corer deployed from the frozen lake surface⁵¹. To prevent eventual internal mixing, we discarded all upper suspended sediments and only kept the compacted sediment for further investigation. Cores were cut into smaller sections to allow transport and storage. All cores were taken to laboratories at the University of Calgary and were stored cold at 5 °C until subsequent subsampling. Cores were split using an adjustable tile saw, cutting only the PVC pipe. The split half was taken into a positive pressure laboratory for DNA subsampling. DNA samples were taken wearing full body suit, mask and sterile gloves; the top 10 mm were removed using two sterile scalpels and samples were taken with a 5 ml sterile disposable syringe (3–4 cm²) and transferred to a 15 ml sterile spin tube. Caution was taken not to cross-contaminate between layers or to sample sediments in contact with the inner side of the PVC pipe. Samples were taken every centimetre in the lowest 1 m of the core (except for Spring Lake, the lowest 2 m), then intervals of 2 cm higher up, and finally samples were taken every 5 cm, and subsequently frozen until analysed. Pollen samples were taken immediately next to the DNA samples, while macrofossil samples were cut from the remaining layer in 1 cm or 2 cm slices. Following sampling, the second intact core halves were visually described and wrapped for transport. All cores were stored at 5 °C before, during and after shipment to the University of Copenhagen (Denmark).

Core logging and scanning. An ITRAX core scanner was used to take high-resolution images and to measure magnetic susceptibility at the Department of Geoscience, Aarhus University. Magnetic susceptibility⁵² was measured every 0.5 cm using a Bartington Instruments MS2 system (Extended Data Fig. 2).

Pollen and macrofossil extraction and identification. Pollen was extracted using a standard protocol³⁰. *Lycopodium* markers were added to determine pollen concentrations⁵³ (see Supplementary Information). Samples were mounted in (2000 cs) silicone oil and pollen including spores were counted using a Leica Laborlux-S microscope at 400× magnification and identified using keys^{30,53,54} as well as reference collections of North American and Arctic pollen housed at the University of Alberta and the Danish Natural History Museum, respectively. Pollen and pteridophyte spores were identified at least to family level and, more typically, to genera. Green algae coenobia of *Pediastrum boryanum* and *Botryococcus* were recorded to track changes in lake trophic status. Pollen influx values were calculated using pollen concentrations divided by the deposition rate (see Supplementary Information). Microfossil diagrams were produced and analysed using PSIMPOLL 4.10 (ref. 31). The sequences were zoned with CONIIC³¹, with a stratigraphy constrained clustering technique using the information statistic as a distance measure. All macrofossils were retrieved using a 100 µm mesh size and were identified but not quantified.

Radiocarbon dating and age-depth modelling. Plant macrofossils identified as terrestrial taxa (or unidentifiable macrofossils with terrestrial characteristics where no preferable material could be identified) were selected for radiocarbon (¹⁴C) dating of the lacustrine sediment. All macrofossils were subjected to a standard acid-base-acid (ABA) chemical pre-treatment at the Oxford Radiocarbon Accelerator Unit (ORAU), following a standard protocol⁵⁵, with appropriate 'known age' (that is, independently dendrochronologically-dated tree-ring) standards run alongside the unknown age plant macrofossil samples⁵⁶. Specifically, this ABA chemical pre-treatment (ORAU laboratory pre-treatment code 'VV') involved successive 1 M HCl (20 min, 80 °C), 0.2 M NaOH (20 min, 80 °C) and 1 M HCl (1 h, 80 °C) washes, with each stage followed by rinsing to neutrality (≥3 times) with ultrapure MilliQ deionised water. The three principal stages of this process (successive ABA washes) are similar across most radiocarbon laboratories and are, respectively, intended to remove: (i) sedimentary- and other carbonate contaminants; (ii) organic (principally humic- and fulvic-) acid contaminants; and (iii) any dissolved atmospheric CO₂ that might have been absorbed during the preceding base wash. Thus, any potential secondary carbon contamination was removed, leaving the samples pure for combustion and graphitisation. Accelerator mass spectrometry (AMS) ¹⁴C dating was subsequently performed on the 2.5 MV HVEE tandem AMS system at ORAU⁵⁷. As is standard practice, measurements were corrected for natural isotopic fractionation by normalizing the data to a standard $\delta^{13}\text{C}$ value of −25‰ VPDB, before reporting as conventional ¹⁴C ages before present (BP, before AD 1950)⁵⁸.

These ¹⁴C data were calibrated with the IntCal13 calibration curve⁵⁹ and modelled using the Bayesian statistical software OxCal v. 4.2 (ref. 60). Poisson process ('P_Sequence') deposition models were applied to each of the Charlie and

Spring Lake sediment profiles⁶¹, with objective 'Outlier' analysis applied to each of the constituent ¹⁴C determinations⁶². The P_Sequence model takes into account the complexity (randomness) of the underlying sedimentation process, and thus provides realistic age-depth models for the sediment profiles on the calibrated radiocarbon (IntCal) timescale. The rigidity of the P_Sequence (the regularity of the sedimentation rate) is determined iteratively within OxCal through a model averaging approach, based upon the likelihood (calibrated ¹⁴C) data included within the model⁶⁰. A prior 'Outlier' probability of 5% was applied to each of the ¹⁴C determinations, because there was no reason, a priori, to believe that any samples were more likely to be statistical outliers than others. All ¹⁴C determinations are provided in Extended Data Table 1; OxCal model coding is provided in the Supplementary Information; and plots of the age-depth models derived for Spring and Charlie Lakes are given in Extended Data Fig. 2.

DNA analysis. All DNA extractions and pre-PCR analyses were performed in the ancient DNA facilities of the Centre for GeoGenetics, Copenhagen. Total genomic DNA was extracted using a modified version of an organic extraction protocol⁶³. We used a lysis buffer containing 68 mM *N*-lauroylsarcosine sodium salt, 50 mM Tris-HCl (pH 8.0), 150 mM NaCl, and 20 mM EDTA (pH 8.0) and, immediately before extraction, 1.5 ml 2-mercaptoethanol and 1.0 ml 1 M DTT were added for each 30 ml lysis buffer. Approximately 2 g of sediment was added, and 3 ml of buffer, together with 170 µg of proteinase K, and vortexed vigorously for 2 × 20 s using a FastPrep-24 at speed 4.0 ms^{−1}. An additional 170 µg of proteinase K was added to each sample and incubated, gently rotating overnight at 37 °C. For removal of inhibitors we used the MOBIO (MO BIO Laboratories, Carlsbad, CA) C2 and C3 buffers following the manufacturer's protocol. The extracts were further purified using phenol-chloroform and concentrated using 30 kDa Amicon Ultra-4 centrifugal filters as described in the Andersen extraction protocol⁶³. Our extraction method was changed from this protocol with the following modifications: no lysis matrix was added due to the minerogenic nature of the samples and the two phenol, one chloroform step was altered, thus both phenol:chloroform:supernatant were added simultaneously in the respective ratio 1:0.5:1, followed by gentle rotation at room temperature for 10 min and spun for 5 min at 3,200g. For dark-coloured extracts, this phenol:chloroform step was repeated. All extracts were quantified using Quant-iT dsDNA HS assay kit (Invitrogen) on a Qubit 2.0 Fluorometer according to the manufacturer's manual. The measured concentrations were used to calculate the total ng DNA extracted per g of sediment (Fig. 2). 32 samples were prepared for shotgun metagenome sequencing⁶⁴ using the NEBNext DNA Library Prep Master Mix Set for 454 (New England BioLabs) following the manufacturer's protocol with the following modifications: (i) all reaction volumes (except for the end repair step) were decreased to half the size as in the protocol, and (ii) all purification steps were performed using the MinElute PCR Purification kit (Qiagen). Metagenome libraries were amplified using AmpliTaq Gold (Applied Biosystems), given 14–20 cycles following and quantified using the 2100 BioAnalyser chip (Agilent). All libraries were purified using Agencourt AMPure XP beads (BeckmanCoulter), quantified on the 2100 BioAnalyzer and pooled equimolarly. All pooled libraries were sequenced on an Illumina HiSeq 2500 platform and treated as single-end reads.

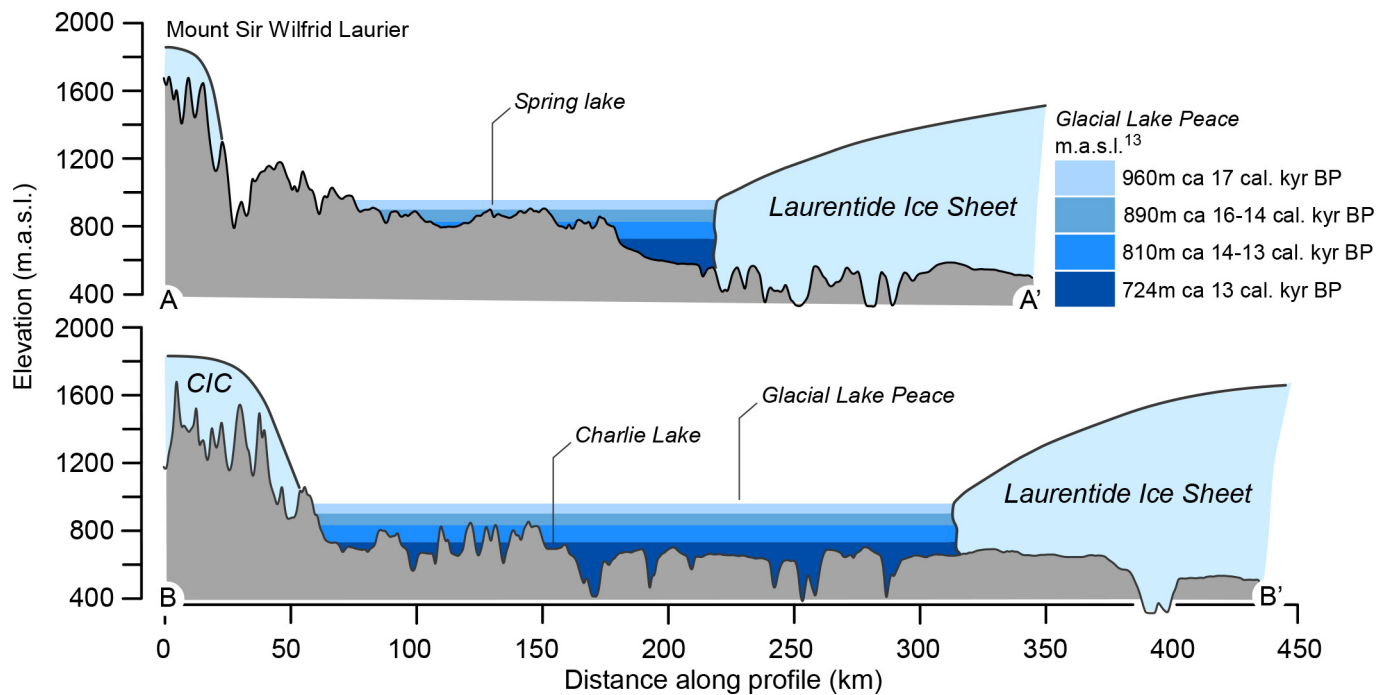
Bioinformatics. Metagenomic reads were demultiplexed and trimmed using AdapterRemoval 1.5 (ref. 65) with a minimum base quality of 30 and minimum length of 30 bp⁶⁶. All reads with poly-A/T tails ≥ 4 were removed from each sample. Low-quality reads and duplicates were removed using String Graph Assembler (SGA)⁶⁷ setting the preprocessing tool dust-threshold = 1, index algorithm = 'ropebwt' and using the SGA filter tool to remove exact and contained duplicates. Each quality-controlled (QC) read was thereafter allowed equal change to map to reference sequences using Bowtie2 version 2.2.4 (ref. 68) (end-to-end alignment and mode -k 50 for example, reads were allowed a total of 500 hits before being parsed). A few reads with more than 500 matches were confirmed by checking that the best blast hit belonged to this taxon, and that alternative hits have lower e-values and alignment scores. We used the full nucleotide database (nt) from GenBank (accessed 4 March 2015), which due to size and downstream handling was divided into 9 consecutive equally sized databases and indexed using Bowtie2-build. All QC checked fastq files were aligned end-to-end using Bowtie2 default settings. Each alignment was merged using SAMtools⁶⁹, sorted according to read identifier and imported to MEGAN v. 10.5 (ref. 70). We performed a lowest common ancestor (LCA) analysis using the built-in algorithm in MEGAN and computed the taxonomic assignments employing the embedded NCBI taxonomic tree (March 2015 version) on reads having 100% matches to a reference sequence. We call this pipeline 'Holi' because it takes a holistic approach because it has no a priori assumption of environment and the read is given an equal chance to align against the nt database containing the vast majority of organismal

sequences (see Supplementary Information). *In silico* testing of 'Holi' sensitivity (see Supplementary Information) revealed 0.1% as a reliable minimum threshold for Viridiplantae taxa. For metazoan reads, which were found to be under-represented in our data, we set this threshold to 3 unique reads in one sample or 3 unique reads in three different samples from the same lake. In addition, we confirmed that each read within the metazoans by checking that the best blast hit belonged to this taxon, and that alternative hits have lower *e*-values and alignment scores⁷¹. We merged all sequences from all blanks and subtracted this from the total data set (instead of pairing for each extract and library build), using lowest taxonomic end nodes. Candidate detection was performed by decreasing the detection threshold in 'Holi' from 0.1% to 0.01% to increase the detection of contaminating plants, and similar for metazoans, we decreased the detection level and subtracted all with 2 or more reads per taxa (see Supplementary Information). We performed a series of *in silico* tests to measure the sensitivity and specificity of our assignment method and to estimate likelihood of false-positives (see Supplementary Information).

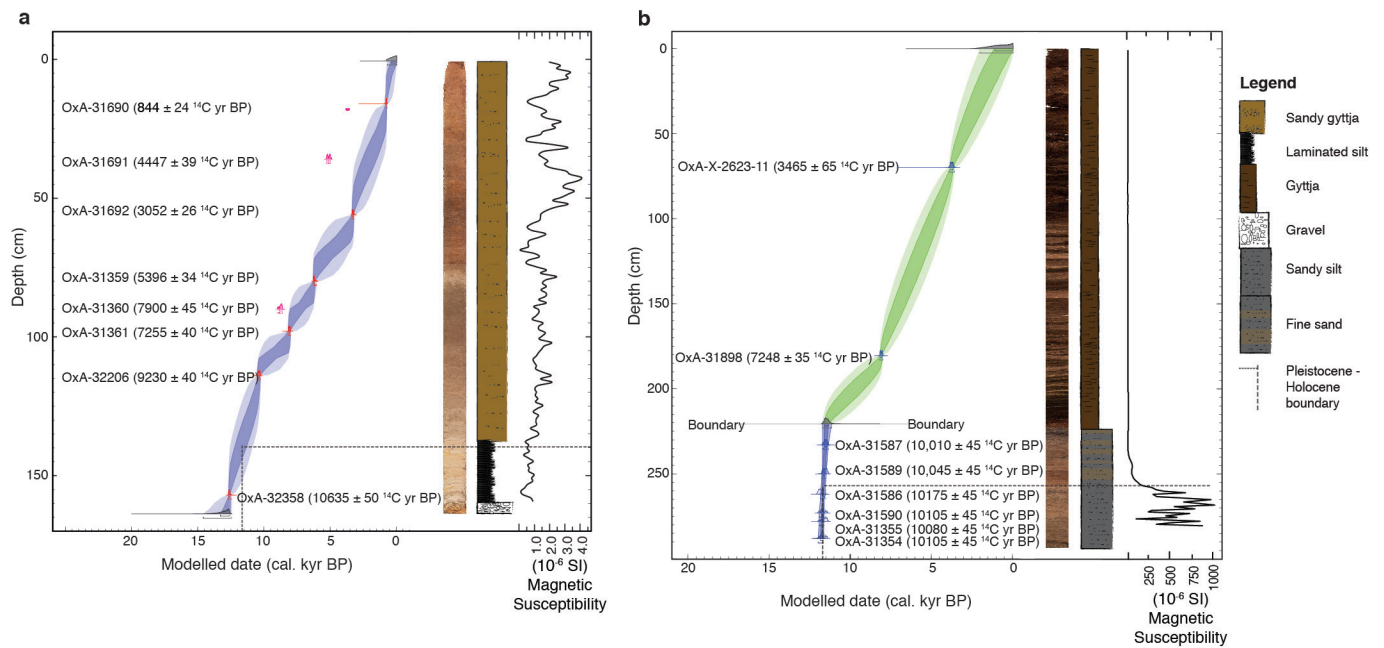
We generated 1,030,354,587 Illumina reads distributed across 32 sediment samples and used the dedicated computational pipeline ('Holi') for handling read de-multiplexing, adaptor trimming, control quality, duplicate and low-complexity read removal (see Supplementary Information). The 257,890,573 reads parsing filters were further aligned against the whole non-redundant nucleotide (nt) sequence database⁷². Hereafter, we used a lowest common ancestor approach⁷⁰ to recover taxonomic information from the 985,818 aligning reads. Plants represented by less than 0.1% of the total reads assigned were discarded to limit false positives resulting from database mis-annotations, PCR and sequencing errors (see Supplementary Information). Given the low number of reads assigned to multicellular, eukaryotic organisms (metazoans), we set a minimal threshold of 3 counts per sample or 1 count in each of three samples. For plants and metazoans this resulted in 511,504 and 2,596 reads assigned at the family or genus levels, respectively. The read counts were then normalized for generating plant and metazoan taxonomic profiles (Extended Data Figs 5 and 6). Taxonomic profiles for reads assigned to bacteria, archaea, fungi and alveolata were also produced (see Supplementary Information).

DNA damage and authenticity. We estimated the DNA damage levels using the MapDamage package 2.0 (ref. 40) for the most abundant organisms (Extended Data Fig. 7b). These represent distinctive sources, which help to account for potential differences between damage accumulated from source to deposition or during deposition. Input SAM files were generated for each sample using Bowtie2 (ref. 68) to align all QC reads from each sample against each reference genome. All aligning sequences were converted to BAM format, sorted and parsed through MapDamage by running the statistical estimation using only the 5'-ends (–forward) for single reads. All frequencies of cytosine to thymine mutations per position from the 5' ends were parsed and the standard deviation was calculated to generate DNA damage models for each lake (Extended Data Fig. 7a and Supplementary Information).

51. Reasoner, M. A. Equipment and procedure improvements for a lightweight, inexpensive, percussion core sampling system. *J. Paleolimnol.* **8**, 273–281 (1993).
52. Sandgren, P. & Snowball, I. *Tracking Environmental Change Using Lake Sediments: Physical and Geochemical Methods* Vol. 2, 217–236 (Kluwer Academic Publishers, 2001).
53. Moore, P. D., Webb, J. A. & Collison, M. E. *Pollen Analysis* 1–216 (Blackwell Scientific Publications, 1991).
54. Beug, H. J. *Leitfaden der Pollenbestimmung* 74–90 (Verlag Dr. Friedrich Pfeil, 2004).
55. Brock, F., Higham, T., Ditchfield, P. & Bronk Ramsey, C. Current pretreatment methods for AMS radiocarbon dating at the Oxford Radiocarbon Accelerator Unit (ORAU). *Radiocarbon* **52**, 103–112 (2010).
56. Staff, R. Wood pretreatment protocols and measurement of tree-ring standards at the Oxford Radiocarbon Accelerator Unit (ORAU). *Radiocarbon* **56**, 709–715 (2014).
57. Bronk Ramsey, C., Higham, T. & Leach, P. Towards high-precision AMS: progress and limitations. *Radiocarbon* **46**, 17–24 (2004).
58. Stuiver, M. & Polach, H. Reporting of ¹⁴C data. *Radiocarbon* **19**, 355–364 (1977).
59. Reimer, P. J., Bard, E., Bayliss, A. & Beck, J. W. IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal. BP. *Radiocarbon* **55**, 1869–1887 (2013).
60. Bronk Ramsey, C. & Lee, S. Recent and planned developments of the program OxCal. *Radiocarbon* **55**, 720–730 (2013).
61. Bronk Ramsey, C. Deposition models for chronological records. *Quat. Sci. Rev.* **27**, 42–60 (2008).
62. Bronk Ramsey, C. Dealing with outliers and offsets in radiocarbon dating. *Radiocarbon* **51**, 1023–1045 (2009).
63. Wales, N., Andersen, K., Cappellini, E., Avila-Arcos, M. C. & Gilbert, M. T. Optimization of DNA recovery and amplification from non-carbonized archaeological remains. *PLoS One* **9**, e86827 (2014).
64. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, <http://dx.doi.org/10.1101/pdb.prot5448> (2010).
65. Lindgreen, S. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res. Notes* **5**, 337 (2012).
66. Schubert, M. *et al.* Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* **13**, 178 (2012).
67. Simpson, J. T. & Durbin, R. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.* **22**, 549–556 (2012).
68. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
69. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
70. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res.* **17**, 377–386 (2007).
71. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
72. NCBI. Nt Database. (<ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nt.gz>) (February 2015).
73. Bronk Ramsey, C. Bayesian analysis of radiocarbon dates. *Radiocarbon* **51**, 337–360 (2009).

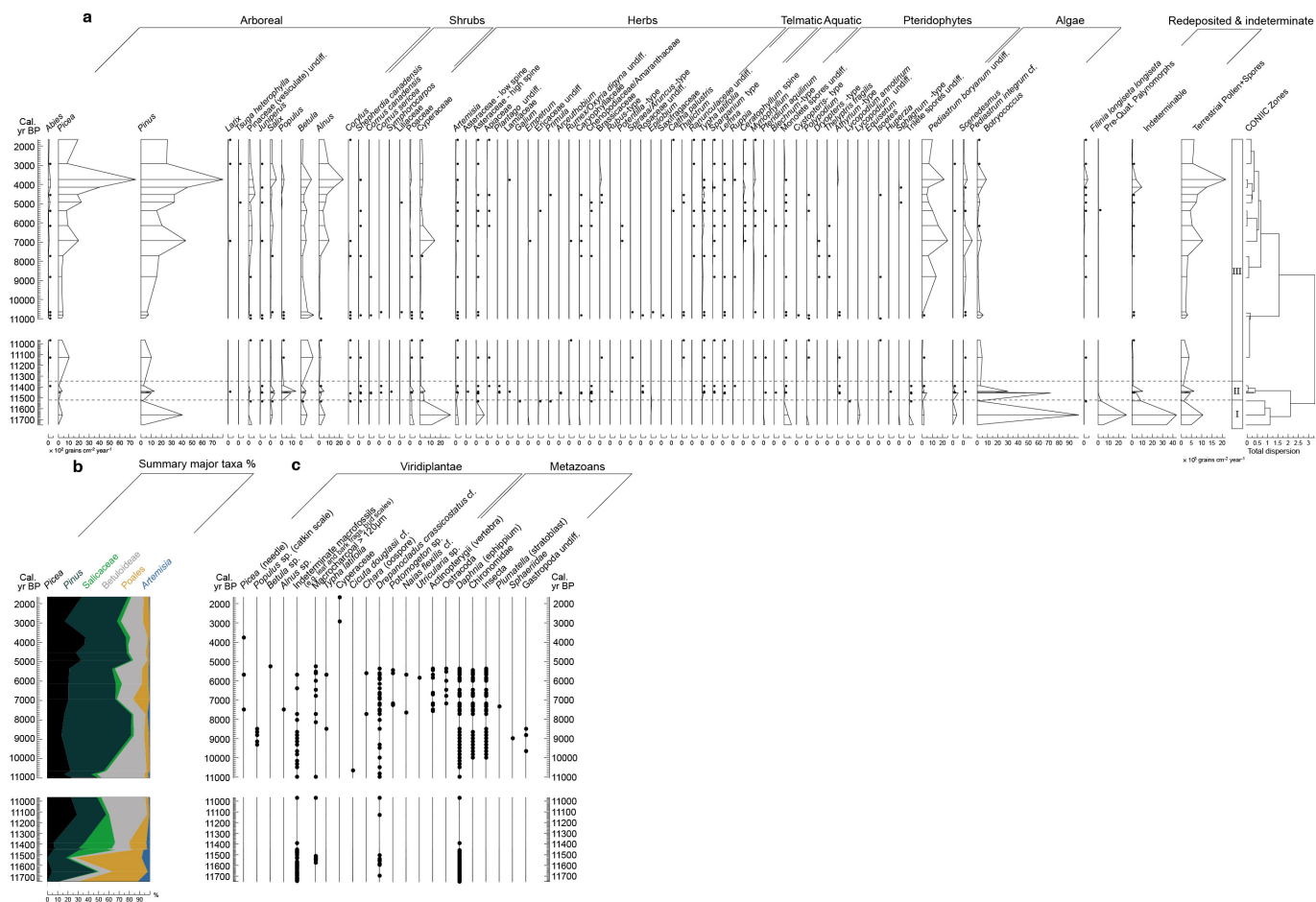


Extended Data Figure 1 | Topographic transects. The red and white lines on Fig. 1b mark topographic transects of Charlie Lake and Spring Lake in relation to the four phases of Glacial Lake Peace¹³. CIC, Cordilleran ice complex; m.a.s.l., metres above sea level.

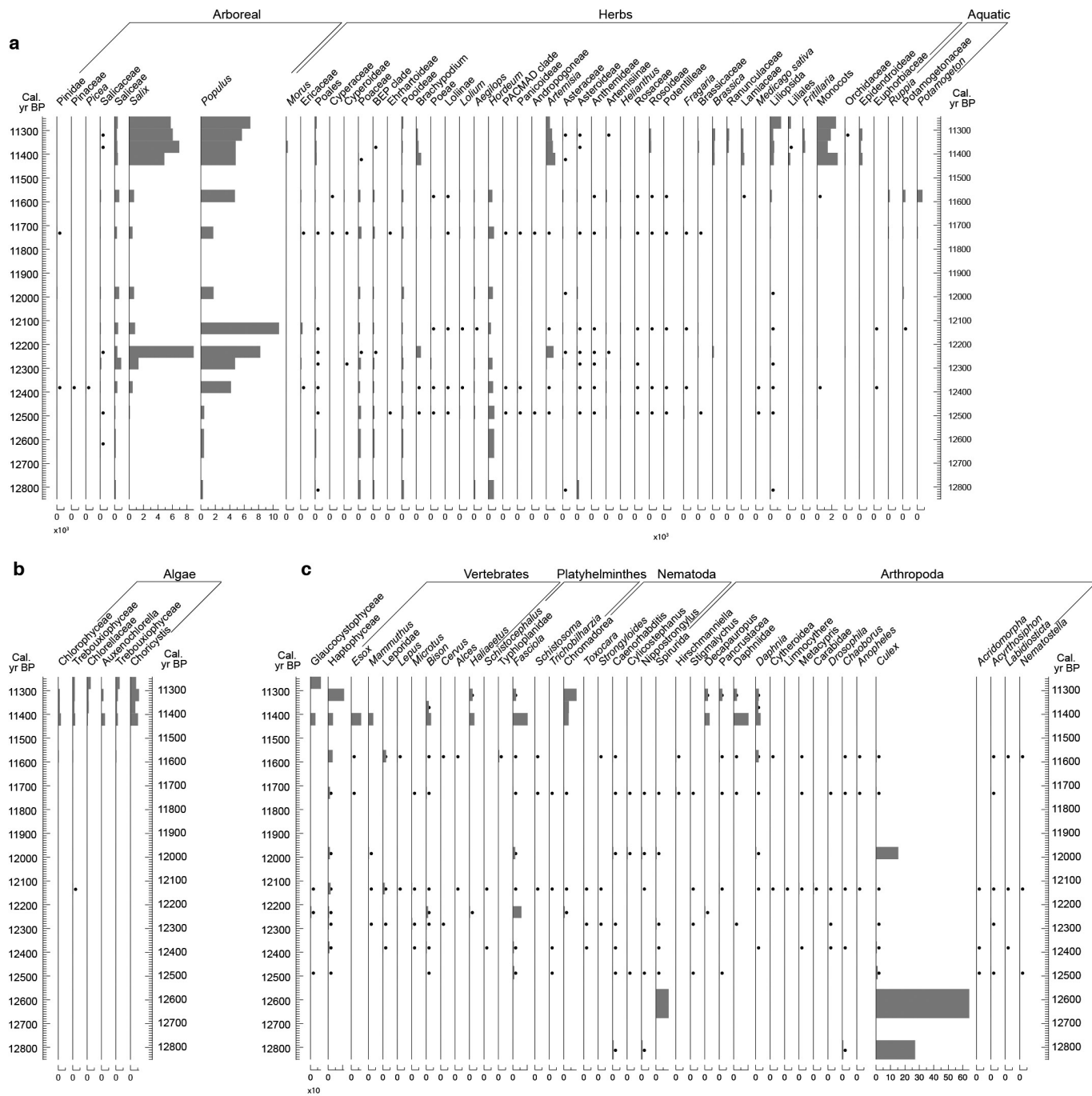


Extended Data Figure 2 | Visual and physical descriptions and age-depth model for the studied lake sediments. a, b, Charlie Lake (a) and Spring Lake (b) span the Pleistocene to Holocene transition (dotted grey line); magnetic susceptibility (continuous black line); and compressed high-resolution images from the ITRAX core scanner and the sedimentary

log are shown. Age-depth models for Charlie Lake (a) and Spring Lake (b) were generated with P_Sequence deposition models in OxCal v. 4.2 using the IntCal13 radiocarbon calibration curve^{57,59,61}. The probability envelopes represent the 68.2% and 95.4% confidence ranges, respectively (see Methods and Supplementary Information).



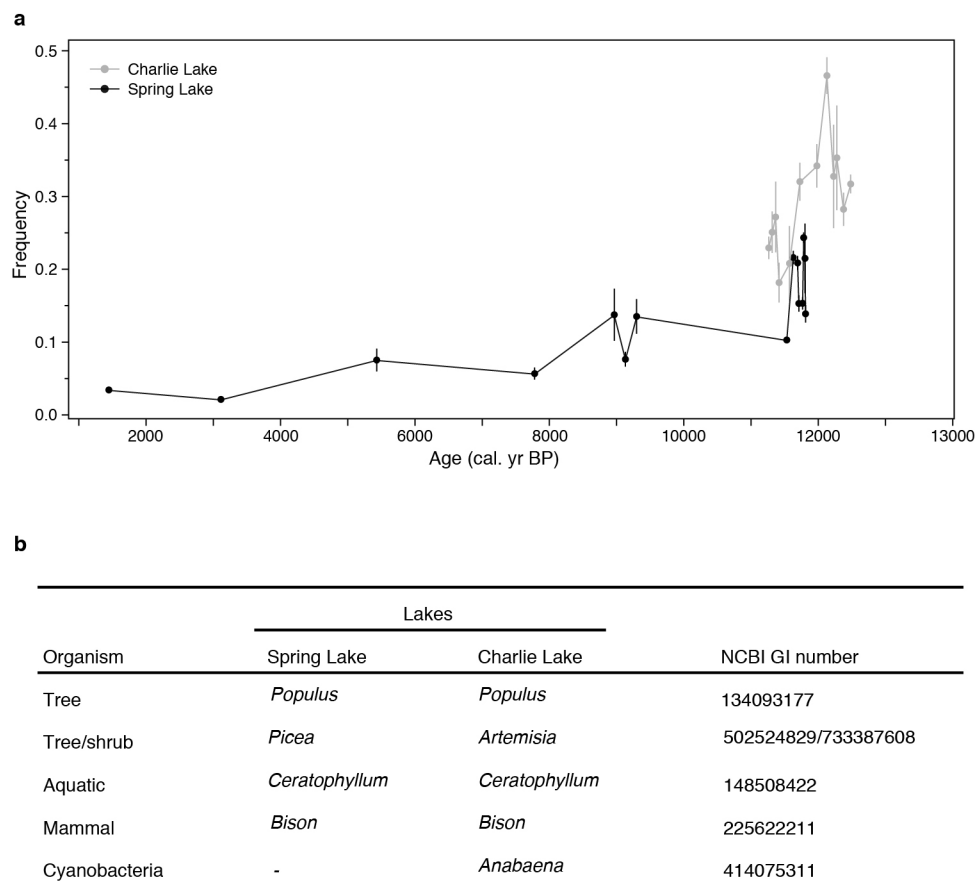
Extended Data Figure 4 | Spring Lake pollen and macrofossil diagrams. **a**, Pollen are presented as influx and bullet points represent taxa with less than 50 grains $\text{cm}^{-2} \text{year}^{-1}$. The diagram was zoned using CONIIC³¹ with a stratigraphically constrained cluster analysis on the information statistic. **b**, Relative proportions of ecologically important taxa. **c**, Macrofossils were identified but not enumerated. Bullet points represent presence.



Extended Data Figure 5 | Charlie Lake DNA diagram. DNA results are presented as normalized counts to allow comparison on the temporal scale for each taxon. All are unique sequences with 100% sequence identity to taxa. Histogram width equals the accumulation period. **a**, Viridiplantae, bullet points represent counts less than 50. **b**, Algae, bullet points represent counts less than 50. **c**, Metazoans, bullet points represent counts equal to 1.



Extended Data Figure 6 | Spring Lake DNA diagram. DNA results are presented as normalized counts to allow comparison on the temporal scale for each taxon. All are unique sequences with 100% sequence identity to taxa. Histogram width equals the accumulation period. **a**, Viridiplantae, bullet points represent counts less than 50. **b**, Algae, bullet points represent counts less than 50. **c**, Metazoans, bullet points represent counts equal to 1.



Extended Data Figure 7 | DNA damage accumulation model. Maximum-likelihood DNA damage rates were estimated from nucleotide misincorporation patterns using MapDamage2.0 (ref. 40). **a**, Each full circle is the mean of cytosine to thymine mutation frequencies at the first position ($n \geq 2$ species) with above 500 reads aligned to reference bars that represent ± 1 s.d. **b**, Table of species used for determining the DNA damage rates.

Extended Data Table 1 | AMS ^{14}C determinations of terrestrial plant macrofossil samples from Charlie and Spring Lakes

Sample ID	AMS Laboratory code	Depth (cm)	Conventional ¹⁴ C		Modeled, calibrated age (cal. BP)		Posterior / Prior Outlier probability (%)	Material dated
			Age (yrs BP ± 1s) unless stated (ref. 82)	δ ¹³ C (‰)				
					68.2% probability range	95.4% probability range		
Charlie Lake								
MWP_30	OxA-31690	16	844 ± 24	-24.9	782 - 730	797 - 690	5 / 5	Unid. terrestrial and charred plants
MWP_31	OxA-31691 *	36	4447 ± 39	-26.0	*	*	(81 / 5) *	Unidentifiable terrestrial plants
MWP_32	OxA-31692	56	3052 ± 26	-27.9	3335 - 3216	3353 - 3180	3 / 5	Unidentifiable terrestrial plants
MWP_20	OxA-31359	80	5396 ± 34	-21.3	6276 - 6188	6289 - 6024	3 / 5	Unidentifiable terrestrial plants
MWP_21	OxA-31360 *	90	7900 ± 45	-24.0	*	*	(70 / 5) *	<i>Cyperus</i> sp. + <i>Carex</i> sp. (seed)
MWP_22	OxA-31361	98	7255 ± 40	-22.8	8156 - 8016	8171 - 7995	4 / 5	<i>Caryx</i> cf. (2 seeds)
MWP_09	OxA-32206	114	9230 ± 40	-26.6	10480 - 10290	10506 - 10257	4 / 5	<i>Potentilla</i> sp., <i>Rumex</i> sp.
MWP_23	OxA-X-2623-15 †*	149	1.03763 ± 0.00595	-25.6	*	*	(100 / 5) *	Unidentifiable terrestrial plants
MWP_10	OxA-31358	157	10635 ± 50	-26.6	12668 - 12573	12715 - 12447	4 / 5	Unidentifiable terrestrial plants
Spring Lake								
MWP_11	OxA-X-2623-11 †	70	3465 ± 65	-22.8	3829 - 3643	3901 - 3568	4 / 5	<i>Picea cf. mariana</i>
MWP_35	OxA-31898	180.5	7248 ± 35	-10.8	8158 - 8014	8170 - 7992	4 / 5	Grass (two ears)
MWP_15	OxA-31587	233	10010 ± 45	-27.1	11604 - 11331	11707 - 11145	3 / 5	Unidentifiable terrestrial plants
MWP_15	OxA-31588	233	10040 ± 45	-28.7	11604 - 11331	11707 - 11145	3 / 5	Unidentifiable terrestrial plants
MWP_16	OxA-31589	250	10045 ± 45	-26.8	11701 - 11556	11746 - 11483	3 / 5	Unidentifiable terrestrial plants
MWP_02	OxA-31586	262	10175 ± 45	-27.8	11735 - 11616	11825 - 11510	6 / 5	Unidentifiable terrestrial plants
MWP_17	OxA-31590	273	10105 ± 45	-26.1	11762 - 11649	11905 - 11615	3 / 5	Unidentifiable terrestrial plants
MWP_18	OxA-31355	278	10080 ± 55	-24.6	11791 - 11662	11925 - 11624	4 / 5	<i>Sphagnaceae</i> , cf. <i>Sphagnum</i>
MWP_03	OxA-31354	288	10105 ± 50	-25.4	11836 - 11673	11978 - 11641	4 / 5	Unidentifiable terrestrial plants

Data were calibrated with the IntCal13 calibration curve⁵⁹ and modelled using the Bayesian statistical software OxCal v. 4.2 (refs. 60, 61, 73).

†Samples that represent 'very small graphite' AMS targets (<0.5 mg C), and so should be treated with caution (and hence the 'OxA-X-' laboratory code prefix).

*Three samples from Charlie Lake produced statistically outlying dates (>50% probability) that were excluded from the final age model.

The antibody aducanumab reduces A β plaques in Alzheimer's disease

Jeff Sevigny^{1*}, Ping Chiao^{1*}, Thierry Bussière^{1*}, Paul H. Weinreb^{1*}, Leslie Williams¹, Marcel Maier², Robert Dunstan¹, Stephen Salloway³, Tianle Chen¹, Yan Ling¹, John O'Gorman¹, Fang Qian¹, Mahin Arastu¹, Mingwei Li¹, Sowmya Chollate¹, Melanie S. Brennan¹, Omar Quintero-Monzon¹, Robert H. Scannevin¹, H. Moore Arnold¹, Thomas Engber¹, Kenneth Rhodes¹, James Ferrero¹, Yaming Hang¹, Alvydas Mikulskis¹, Jan Grimm², Christoph Hock^{2,4}, Roger M. Nitsch^{2,4§} & Alfred Sandrock^{1§}

Alzheimer's disease (AD) is characterized by deposition of amyloid- β (A β) plaques and neurofibrillary tangles in the brain, accompanied by synaptic dysfunction and neurodegeneration. Antibody-based immunotherapy against A β to trigger its clearance or mitigate its neurotoxicity has so far been unsuccessful. Here we report the generation of aducanumab, a human monoclonal antibody that selectively targets aggregated A β . In a transgenic mouse model of AD, aducanumab is shown to enter the brain, bind parenchymal A β , and reduce soluble and insoluble A β in a dose-dependent manner. In patients with prodromal or mild AD, one year of monthly intravenous infusions of aducanumab reduces brain A β in a dose- and time-dependent manner. This is accompanied by a slowing of clinical decline measured by Clinical Dementia Rating—Sum of Boxes and Mini Mental State Examination scores. The main safety and tolerability findings are amyloid-related imaging abnormalities. These results justify further development of aducanumab for the treatment of AD. Should the slowing of clinical decline be confirmed in ongoing phase 3 clinical trials, it would provide compelling support for the amyloid hypothesis.

The amyloid hypothesis posits that A β -related toxicity is the primary cause of synaptic dysfunction and subsequent neurodegeneration that underlies the progression characteristic of AD^{1,2}. Genetic, neuropathological, and cell biological evidence strongly suggest that targeting A β could be beneficial for patients with AD^{3,4}. So far, attempts at therapeutically targeting A β have not been successful^{5–7}, casting doubt on the validity of the amyloid hypothesis. However, the lack of success may have been due to the inability of the antibodies to adequately engage their target or the proper target in the brain, or selecting the wrong patient population.

We describe the development of an antibody-based immunotherapeutic approach by selecting human B-cell clones triggered by neo-epitopes present in pathological A β aggregates. The screening of libraries of human memory B cells for reactivity against aggregated A β led to molecular cloning, sequencing, and recombinant expression of aducanumab (BIIB037), a human monoclonal antibody that selectively reacts with A β aggregates, including soluble oligomers and insoluble fibrils. In preclinical studies, we show that an analogue of aducanumab is capable of crossing the blood–brain barrier, engaging its target, and clearing A β from plaque-bearing transgenic mouse brains. These results prompted the start of clinical trials⁸.

We report here interim results from a double-blind, placebo-controlled phase 1b randomized trial (PRIME; ClinicalTrials.gov identifier NCT01677572) designed to investigate the safety, tolerability, pharmacokinetics, and pharmacodynamics of monthly infusions of aducanumab in patients with prodromal or mild AD with brain A β pathology confirmed by molecular positron emission tomography (PET) imaging. Together, our data support further development of aducanumab as an A β -removing, disease-modifying therapy for AD.

Removal of brain A β plaques in patients with AD

In the PRIME study, 165 patients were randomized and treated between October 2012 and January 2014 at 33 sites in the United States. Patients with a clinical diagnosis of prodromal or mild AD and visually positive A β PET scan⁹ were given monthly intravenous infusions of placebo or aducanumab at doses of 1, 3, 6 or 10 mg kg^{−1} for one year. Of these patients, 125 completed and 40 discontinued treatment, most commonly due to adverse events (20 patients) and withdrawal of consent (14 patients): 25% of the placebo group discontinued compared with 23%, 19%, 17%, and 38% of the 1, 3, 6 and 10 mg kg^{−1} aducanumab dose groups, respectively (Extended Data Fig. 1). Baseline characteristics, including cognitive measures, were generally well-balanced across the groups, although the 1 mg kg^{−1} dose group included a higher proportion of patients with mild AD, and the aducanumab treatment groups tended to have a higher Clinical Dementia Rating—Sum of Boxes (CDR-SB) score (Table 1).

Treatment with aducanumab reduced brain A β plaques as measured by florbetapir PET imaging in a dose- and time-dependent fashion (Fig. 1, 2a). The mean PET standard uptake value ratio (SUVR) composite score at baseline was 1.44. After 54 weeks of treatment, this had decreased significantly ($P < 0.001$) in the 3, 6 and 10 mg kg^{−1} dose groups; whereas change for the placebo group was minimal (Fig. 2a, Extended Data Table 1). In the 10 mg kg^{−1} dose group, the SUVR composite score was 1.16 after 54 weeks of treatment, a value near the purported quantitative cut-point of 1.10 that discriminates between positive and negative scans (Fig. 2b)¹⁰. The adjusted mean changes in SUVR composite scores in the 6 and 10 mg kg^{−1} groups treated for 26 weeks were similar in magnitude to the dose group below (3 and 6 mg kg^{−1}, respectively) treated for 54 weeks (Fig. 2a). Reductions in amyloid PET SUVR composite score in aducanumab-treated patients

¹Biogen, Cambridge, Massachusetts 02142, USA. ²Neurimmune, Schlieren-Zürich 8952, Switzerland. ³Butler Hospital, Providence, Rhode Island 02906, USA. ⁴Institute for Regenerative Medicine, University of Zurich, Zurich 8952, Switzerland.

*These authors contributed equally to this work.

§These authors jointly supervised this work.

Table 1 | Baseline characteristics

Characteristic		Aducanumab					Total (n = 165)*
		Placebo (n = 40)	1 mg kg ⁻¹ (n = 31)	3 mg kg ⁻¹ (n = 32)	6 mg kg ⁻¹ (n = 30)	10 mg kg ⁻¹ (n = 32)	
Years of age (mean ± s.d.)		72.8 ± 7.2	72.6 ± 7.8	70.5 ± 8.2	73.3 ± 9.3	73.7 ± 8.3	72.6 ± 8.1
Female sex (n (%))		23 (58)	13 (42)	17 (53)	15 (50)	15 (47)	83 (50)
ApoE ε4 (n (%))	Carriers	26 (65)	19 (61)	21 (66)	21 (70)	20 (63)	107 (65)
	Non-carriers	14 (35)	12 (39)	11 (34)	9 (30)	12 (38)	58 (35)
Clinical stage (n (%))	Prodromal	19 (48)	10 (32)	14 (44)	12 (40)	13 (41)	68 (41)
	Mild	21 (53)	21 (68)	18 (56)	18 (60)	19 (59)	97 (59)
MMSE (mean ± s.d.)		24.7 ± 3.6	23.6 ± 3.3	23.2 ± 4.2	24.4 ± 2.9	24.8 ± 3.1	24.2 ± 3.5
Global CDR (n (%))	0.5	34 (85)	22 (71)	22 (69)	25 (83)	24 (75)	127 (77)
	1	6 (15)	9 (29)	10 (31)	5 (17)	8 (25)	38 (23)
CDR-SB (mean ± s.d.)		2.66 ± 1.50	3.40 ± 1.76	3.50 ± 2.06	3.32 ± 1.54	3.14 ± 1.71	3.18 ± 1.72
FCSRT sum of free recall score (mean ± s.d.)		15.2 ± 8.5	13.2 ± 9.0	13.8 ± 8.0	14.4 ± 8.3	14.6 ± 8.3	14.3 ± 8.3
PET SUVR composite score (mean ± s.d.)		1.44 ± 0.17	1.44 ± 0.15	1.46 ± 0.15	1.43 ± 0.20	1.44 ± 0.19	1.44 ± 0.17
AD medications use† (n (%))		24 (60)	19 (61)	28 (88)	20 (67)	17 (53)	108 (65)

Percentages are rounded to the nearest integer. AD, Alzheimer's disease; ApoE ε4, apolipoprotein E ε4 allele; CDR, Clinical Dementia Rating; CDR-SB, Clinical Dementia Rating—Sum of Boxes; FCSRT, Free and Cued Selective Reminding Test; MMSE, Mini-Mental State Examination; PET, positron emission tomography; SD, standard deviation; SUVR, standard uptake value ratio.

*Number of patients dosed.

†Cholinesterase inhibitors and/or memantine.

were similar in patients with mild and prodromal AD, and apolipoprotein E (ApoE) ε4 carriers and non-carriers (Extended Data Fig. 2a, b). Pre-specified regional analyses of SUVR changes demonstrated statistically significant dose-dependent reductions in all brain regions, except for the pons and sub-cortical white matter, two areas in which Aβ plaques are not expected to accumulate (Extended Data Fig. 3).

Effect on clinical measures

Clinical assessments were exploratory as the study was not powered to detect clinical change. The test of dose response was the pre-specified primary analysis for the clinical assessments. Analysis of change from baseline on the CDR-SB (adjusted for baseline CDR-SB and ApoE ε4 status) demonstrated dose-dependent slowing of clinical progression with aducanumab treatment at one year (dose-response, $P < 0.05$), with the greatest slowing for 10 mg kg⁻¹ ($P < 0.05$ versus placebo) (Fig. 3a, Extended Data Table 1). Sensitivity analysis using a mixed model for repeated measures (MMRM) also showed a trend for slowing of decline on the CDR-SB at one year ($P = 0.07$ with 10 mg kg⁻¹ aducanumab versus placebo). A dose-dependent slowing of clinical progression on the Mini Mental State Examination (MMSE) with aducanumab treatment was also observed at one year (dose-response, $P < 0.05$), with the greatest effects at 3 and 10 mg kg⁻¹ aducanumab ($P < 0.05$ versus placebo) (Fig. 3b, Extended Data Table 1). On sensitivity analysis using MMRM, the greatest difference was retained for 10 mg kg⁻¹ aducanumab ($P < 0.05$ versus placebo), with a smaller difference at 3 mg kg⁻¹ ($P = 0.10$ versus placebo). No changes from baseline after one year were found on the composite neuropsychological test battery (NTB) or the Free and Cued Selective Reminding Test (FCSRT) free recall (Extended Data Table 1), but skewed non-normal (floor) effects at baseline were observed. The floor effects on the NTB were seen in the individual tests; specifically, in the two most clinically relevant components given the stage of the population enrolled: Wechsler Memory Scale-Fourth Edition Verbal Paired Associates II (WMS-IV VPA II) and Rey Auditory Verbal Learning Test (RAVLT) delayed recall of the NTB memory domain.

Safety and tolerability

The most common adverse effects were amyloid-related imaging abnormalities (ARIA), headache, urinary tract infection, and upper respiratory tract infection (Table 2). Using the most specific descrip-

tion of ARIA by magnetic resonance imaging (MRI), ARIA-vasogenic oedema (ARIA-E) abnormalities occurred in no patients receiving placebo compared with 1 (3%), 2 (6%), 11 (37%), and 13 (41%) patients receiving 1, 3, 6 and 10 mg kg⁻¹ aducanumab, respectively (Extended Data Table 2). ARIA-E was generally observed early in the course of treatment, MRI findings typically resolved within 4–12 weeks, and of the 27 patients who developed ARIA-E, 15 (56%) continued treatment (Supplementary Information). All cases of symptomatic ARIA were

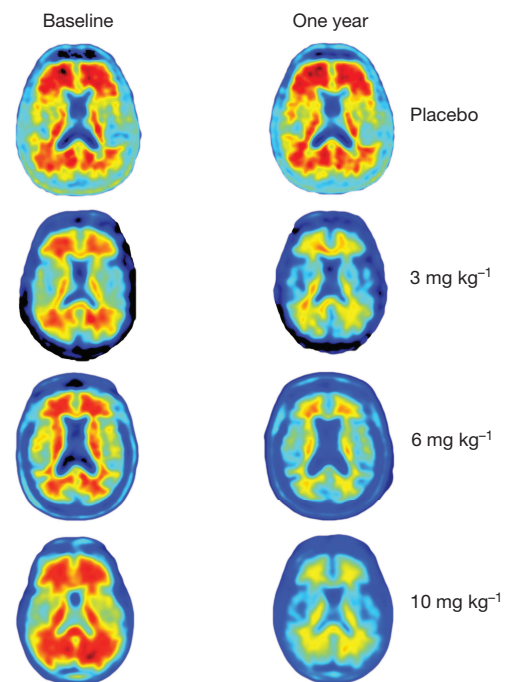


Figure 1 | Amyloid plaque reduction with aducanumab: example amyloid PET images at baseline and week 54. Individuals were chosen based on visual impression and SUVR change relative to average one-year response for each treatment group ($n = 40, 32, 30$ and 32 , respectively). Axial slice shows anatomical regions in posterior brain putatively related to AD pathology. SUVR, standard uptake value ratio.

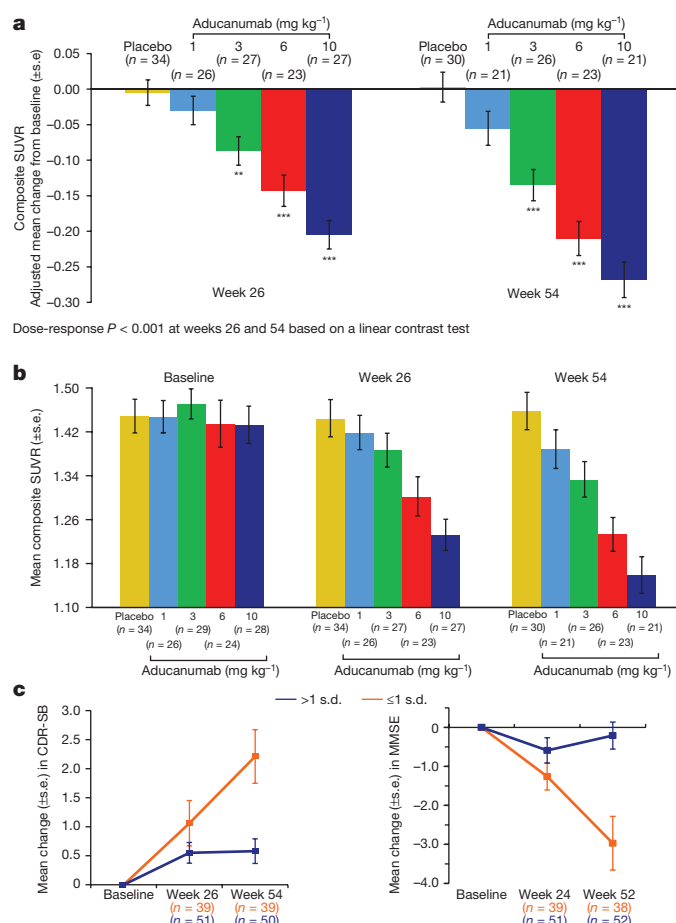


Figure 2 | Amyloid plaque reduction with aducanumab. a–c, Change from baseline (a, analyses using ANCOVA), SUVR values (b), and categorization of change in amyloid PET (c) at week 54 and associated change from baseline CDR-SB and MMSE in aducanumab-treated patients (post hoc analysis). Categorization of amyloid PET at week 54 based on s.d. of change from baseline in placebo-treated patients. ** $P < 0.01$; *** $P < 0.001$ versus placebo; two-sided tests with no adjustments for multiple comparisons. Mean \pm s.e. ANCOVA, analysis of covariance; CDR-SB, Clinical Dementia Rating—Sum of Boxes; MMSE, Mini Mental State Examination; SUVR, standard uptake value ratio.

required to be reported as medically important serious adverse effects. No patients were hospitalised for ARIA. The only serious adverse effects (by preferred term) that occurred in more than one patient in any treatment group were ARIA (0, 1 (3%), 1 (3%), 4 (13%), and 5 (16%) of patients receiving placebo, and 1, 3, 6 and 10 mg kg⁻¹ aducanumab, respectively) and superficial siderosis of the central nervous system (0, 1 (3%), 0, 2 (7%), and 3 (9%) of patients receiving placebo and 1, 3, 6 and 10 mg kg⁻¹ aducanumab, respectively). Owing to the requirement for repeated MRI assessments of those patients who developed ARIA, these individuals were partially unblinded to treatment. Other adverse effects and serious adverse effects were consistent with the patient population. There were no drug-related deaths (Supplementary Information).

Pharmacokinetics

The pharmacokinetics of aducanumab (maximum concentration (C_{max}) and cumulative area under the concentration curve (AUC)) were linear across the dose range in patients who received all 14 planned doses (Extended Data Table 3). The median plasma half-life was 21 days. In total, 3 of 118 evaluable patients (3%) in the combined aducanumab groups developed treatment-emergent anti-aducanumab antibodies within the first year of treatment. Antibody responses were

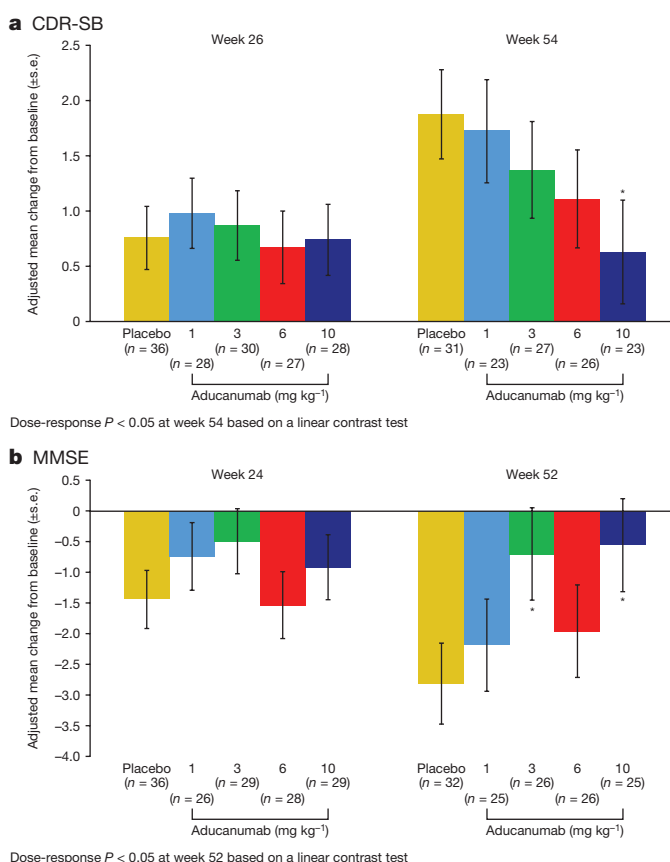


Figure 3 | Aducanumab effect (change from baseline) on CDR-SB and MMSE. a, b, Aducanumab effect on CDR-SB (a) and MMSE (b).

* $P < 0.05$ versus placebo; two-sided tests with no adjustments for multiple comparisons. CDR-SB and MMSE were exploratory endpoints. Adjusted mean \pm s.e. Analyses using ANCOVA. CDR-SB, Clinical Dementia Rating—Sum of Boxes; MMSE, Mini Mental State Examination.

transient, with minimal titres, and had no apparent effect on aducanumab pharmacokinetics or safety.

Brain penetration and binding to A β plaques

In the preclinical studies which preceded PRIME, systemically administered aducanumab (single dose, 30 mg kg⁻¹ intraperitoneally (i.p.)) bound to diffuse and compact A β plaques in the brains of 22-month-old female Tg2576 transgenic mice ('Target engagement study'; Extended Data Fig. 4a–d). C_{max} in plasma was 181 μ g ml⁻¹, with a terminal elimination half-life ($t_{1/2}$) of 2.5 days. The C_{max} in brain was 1,062 ng g⁻¹ of tissue, and approximately 400–500 ng g⁻¹ of drug was measured 3 weeks after dosing, suggesting long-term retention. Consequently, the brain:plasma AUC ratio of 1.3% was higher than the 0.1% frequently reported for systemically administered antibodies^{11,12}.

Administration of a single dose of aducanumab did not affect plasma (Extended Data Fig. 4b) or brain (data not shown) A β concentrations, consistent with the observation that aducanumab does not bind to soluble A β monomers. In contrast, the murine bapineuzumab precursor antibody 3D6, which binds to A β monomers, triggered a transient plasma A β spike (Extended Data Fig. 4b). Similarly, plasma A β concentrations were stable after repeated dosing with aducanumab in the PRIME study (data not shown). Within 24 h of dosing, aducanumab bound to parenchymal brain A β with a spatial pattern essentially superimposable with *ex vivo* pan-A β antibody staining, confirming that aducanumab binds all morphological types of brain A β plaques *in vivo*, including diffuse A β deposits and compact A β plaques (Extended Data Fig. 4c, d). Aducanumab binding to A β deposited in cerebral amyloid angiopathy (CAA) lesions within brain blood vessel walls was less

Table 2 | Summary of adverse events and most common adverse events

Adverse event (n (%))	Placebo (n = 40)	Aducanumab			
		1 mg kg ⁻¹ (n = 31)	3 mg kg ⁻¹ (n = 32)	6 mg kg ⁻¹ (n = 30)	10 mg kg ⁻¹ (n = 32)
Any adverse event	39 (98)	28 (90)	27 (84)	28 (93)	29 (91)
Serious event	15 (38)	3 (10)	4 (13)	4 (13)	12 (38)
Discontinuing treatment due to an adverse event	4 (10)	3 (10)	2 (6)	3 (10)	10 (31)
Common adverse events					
ARIA	2 (5)	2 (6)	4 (13)	11 (37)	15 (47)
Headache	2 (5)	5 (16)	4 (13)	8 (27)	8 (25)
Urinary tract infection	4 (10)	3 (10)	2 (6)	4 (13)	5 (16)
Upper respiratory tract infection	6 (15)	2 (6)	2 (6)	2 (7)	6 (19)
Diarrhoea	3 (8)	0	6 (19)	1 (3)	3 (9)
Arthralgia	2 (5)	0	6 (19)	2 (7)	1 (3)
Fall	8 (20)	3 (10)	2 (6)	2 (7)	2 (6)
Superficial siderosis of CNS	0	2 (6)	1 (3)	2 (7)	4 (13)
Constipation	0	3 (10)	1 (3)	1 (3)	3 (9)
Nausea	2 (5)	2 (6)	5 (16)	0	1 (3)
Anxiety	4 (10)	4 (13)	1 (3)	1 (3)	1 (3)
Nasopharyngitis	0	1 (3)	5 (16)	0	1 (3)
Cough	2 (5)	3 (10)	1 (3)	0	1 (3)
Alanine aminotransferase increased	0	3 (10)	0	1 (3)	0
Aspartate aminotransferase increased	0	3 (10)	0	0	1 (3)

Common adverse events are those with an incidence of $\geq 10\%$ in any aducanumab treatment group. Incidence of ARIA based on adverse event reporting. Adverse events of ARIA-E (oedema) and ARIA-H (micro-haemorrhage) are both coded to the MedDRA preferred term of amyloid-related imaging abnormalities, and ARIA-H (superficial siderosis) codes to the MedDRA preferred term of superficial siderosis of the CNS. ARIA, amyloid-related imaging abnormalities; CNS, central nervous system; MedDRA, Medical Dictionary for Regulatory Activities.

prominent than parenchymal A β binding, when compared with the total amount of A β (Extended Data Fig. 4c, d).

Reduction of brain A β in transgenic mice

Exposure in plasma and brain correlated linearly with dose after chronic dosing in plaque-bearing transgenic mice (Extended Data Fig. 5) (Supplementary Information). ^{ch}aducanumab, a murine IgG2a/ κ chimaeric analogue, dose-dependently reduced A β measured in brain homogenates by up to 50% relative to the vehicle control in the diethylamine (DEA) fraction that extracted soluble monomeric and oligomeric forms of A β_{40} and A β_{42} , and in the guanidine hydrochloride (GuHCl) fraction that extracted insoluble A β fibrils (Fig. 4a, b).

Quantitative 6E10 immunohistochemistry showed significant reductions in all forms of A β deposits by up to 70% (Fig. 4c, d). Thioflavin S (ThioS) staining of compact A β plaques showed dose-dependent and statistically significant reductions in the cortex and hippocampus by up to 63% (Fig. 4c, d). Quantitative histology indicated that ^{ch}aducanumab significantly reduced the number of plaques of all sizes, including plaques $>500\mu\text{m}^2$ and plaques $<125\mu\text{m}^2$ (Extended Data Fig. 6a–c). Quantification of ThioS-positive vascular and parenchymal A β plaques separately showed that ^{ch}aducanumab did not affect vascular A β in either cortex or hippocampus (Fig. 4e–h).

To identify the mechanism of A β clearance, we analysed the involvement of microglia which are known to display enhanced phagocytic activities through binding to the Fc region of an antibody^{13,14}. ^{ch}aducanumab significantly increased recruitment of Iba-1-positive microglia to A β plaques, suggesting Fc γ R-mediated phagocytosis of antibody–A β complexes as a possible clearance mechanism (Extended Data Fig. 7a–c and Supplementary Information).

Biochemical characterization

The apparent affinities of aducanumab and ^{ch}aducanumab for aggregated A β_{42} , with half maximal effective concentration (EC₅₀) values of 0.1 nM, were comparable to 3D6 (ref. 13) (Fig. 5a). Neither aducanumab nor ^{ch}aducanumab bound monomeric soluble A β_{40} at concentrations

up to 1 μM , indicating $>10,000$ -fold selectivity for aggregated A β over monomer, whereas 3D6 bound soluble A β_{40} with an EC₅₀ of 1 nM (Fig. 5b). In contrast to 3D6, which immunoprecipitated both monomeric and aggregated A β , ^{ch}aducanumab bound soluble A β_{42} oligomers and insoluble A β_{42} fibrils prepared *in vitro*, but not A β_{42} monomers (Fig. 5c). Histological staining of autopsy tissue from patients with AD or aged amyloid precursor protein (APP) transgenic mice confirmed binding of aducanumab to bona fide human A β fibrils (Fig. 5d, e).

Discussion

The PRIME study shows that aducanumab penetrates the brain and decreases A β in patients with AD in a time- and dose-dependent manner. Within 54 weeks of treatment, 3, 6 and 10 mg kg⁻¹ doses of aducanumab significantly decreased the amyloid PET SUVR. Patients receiving placebo showed virtually no change in their mean PET SUVR composite scores over one year, indicating that A β pathology had already reached an asymptote of accumulation. Considering that it may have taken up to 20 years for A β to have accumulated to the levels in these patients at study entry¹⁵, the observed kinetics of A β removal within a 12-month time period appears encouraging for a disease-modifying treatment for patients with AD.

The cognitive results for CDR-SB and MMSE provide support for the clinical hypothesis that reduction of brain A β confers a clinical benefit. Post hoc analysis showed that those aducanumab-treated patients who had decreased SUVR scores >1 standard deviation unit relative to placebo-treated patients after one year of treatment experienced a stabilization of clinical decline on both CDR-SB and MMSE scores; whereas, those patients with a smaller or no decrease experienced clinical decline similar to placebo patients (Fig. 2c). The apparent clinical benefit observed in PRIME could also be explained by the binding of aducanumab to oligomeric forms of A β , which would not be directly detected by PET imaging. The reductions in SUVR scores may be surrogates for reductions in toxic soluble A β oligomers which may have had a more functionally relevant impact on cognition. Whereas significant A β reduction was detectable by 6 months, clinical effects were not

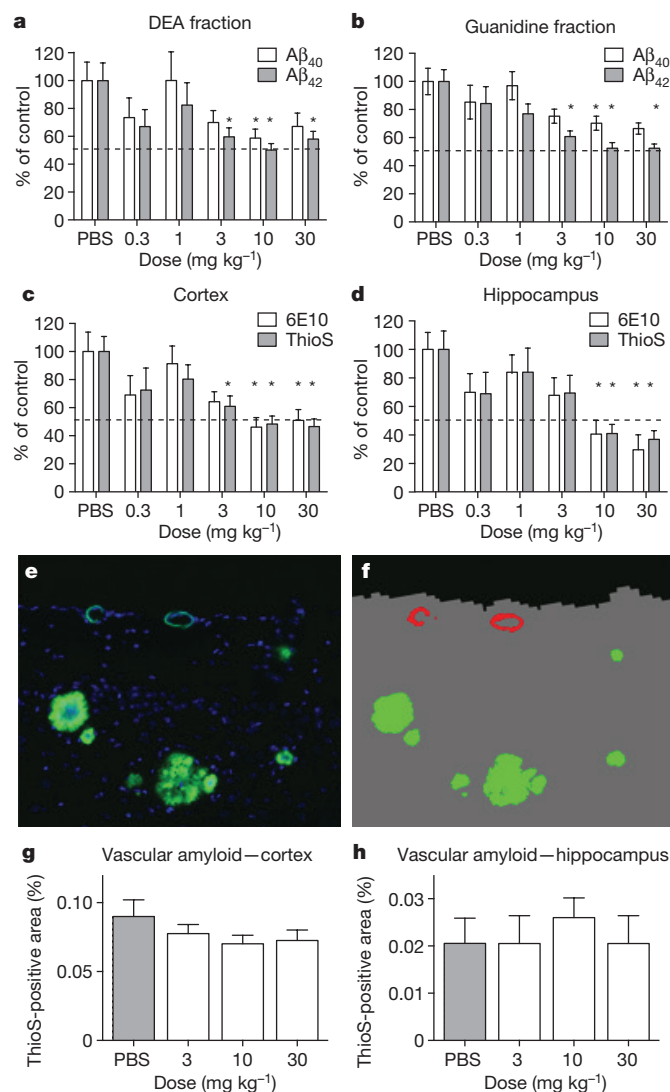


Figure 4 | Reduction of amyloid burden following weekly dosing with *ch*aducanumab in 9.5- to 15.5-month-old Tg2576 transgenic mice.

a, b, $A\beta_{40}$ and $A\beta_{42}$ levels in soluble DEA (**a**) and insoluble GuHCl (**b**) brain fractions. **c, d**, Total brain $A\beta$ (6E10) and compact amyloid plaques (ThioS) in cortex (**c**) and hippocampus (**d**) (mean \pm s.e.; $n = 20$ –55; dotted line 50% reduction; * $P < 0.05$ versus control). **e–h**, ThioS staining of amyloid deposits (**e**) and Visiopharm software (**f**) differentiated parenchymal deposits (green) from vascular deposits (red) (representative pictures 10 \times magnification), and quantified area of vascular amyloid (**g, h**; mean \pm s.e.; $n = 20$ –24).

apparent until one year. Given that clearance of $A\beta$ could be followed by recovery of neuronal function, a lag between reduction of $A\beta$ burden and slowing of disease progression is not altogether surprising.

The main safety finding, ARIA-E, was dose-dependent and more common in ApoE $\epsilon 4$ carriers, consistent with findings with other anti- $A\beta$ monoclonal antibodies^{7,16,17}. Although the underlying cause of ARIA is not well understood, it is likely that the MRI signal of ARIA is due to increased extracellular fluid. This may be a result of underlying CAA, changes in perivascular clearance and vascular integrity, or local inflammatory processes associated with $A\beta$ -targeting therapies^{17–20} (see Supplementary Information for further discussion).

Study limitations of the PRIME phase 1b study included staggered parallel-group design, small sample sizes, limited region (USA only), and possible partial unblinding due to ARIA-E. Measures were taken to maintain blinding to adverse effects: raters of given tests were not permitted to perform other clinical assessments, and were blinded to other assessments (for example, MMSE and CDR raters were required

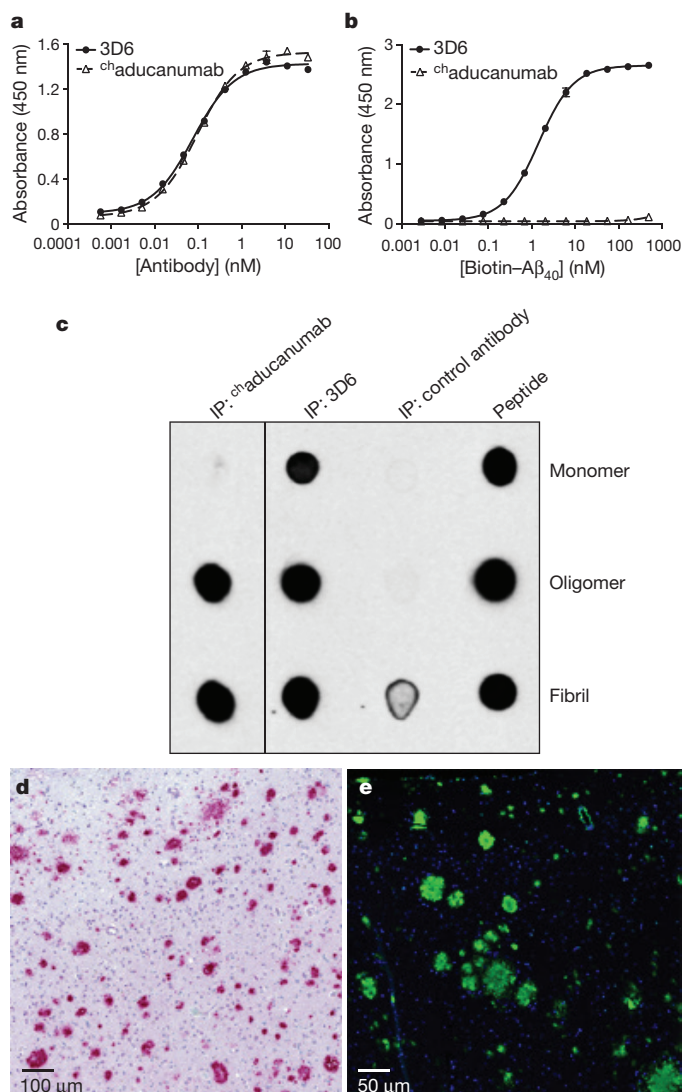


Figure 5 | Aducanumab binds selectively to insoluble fibrillar and soluble oligomeric $A\beta$ aggregates. **a**, Binding of *ch*aducanumab or 3D6 to immobilized fibrillar $A\beta_{42}$. Mean \pm s.d., in triplicate. **b**, Capture of soluble monomeric $A\beta_{40}$ with immobilized *ch*aducanumab or 3D6. Mean \pm s.d., in triplicate. **c**, Dot blots of $A\beta_{42}$ monomer, soluble oligomers, or insoluble fibrils immunoprecipitated with *ch*aducanumab, 3D6, or irrelevant antibody control. Equivalent concentrations confirmed by direct dot blotting (Peptide). **d, e**, Immunostaining of $A\beta$ in autopsy brain tissue from a patient with AD with *ch*aducanumab (0.2 μ g ml⁻¹) (**d**) and 22-month-old Tg2576 transgenic mouse brain tissue with aducanumab (60 ng ml⁻¹) (**e**).

to be different and neither were permitted to perform other study assessments). Post hoc analyses of change from baseline PET SUVR composite score and cognition by presence/absence of ARIA suggested no apparent difference in treatment effect when comparing patients with and without ARIA-E (Extended Data Table 4). There was overlap in Arms 1–3 (aducanumab 1 and 3 mg kg⁻¹, placebo) and Arms 4 and 5 (aducanumab 10 mg kg⁻¹, placebo) but Arms 6 and 7 (aducanumab 6 mg kg⁻¹, placebo) were initiated after enrolment in Arms 1–5 was complete. This was a small study designed for assessment of safety and tolerability, and for detecting a pharmacological effect on brain $A\beta$ levels measured by PET imaging. The trial was not powered for the exploratory clinical endpoints, thus the clinical cognitive results should be interpreted with caution. Primary analyses were based on observed data with no imputation for missing values, nominal P values were presented with no adjustments for multiple comparisons, and they were supported by sensitivity analyses using a MMRM.

The initiation of the PRIME study and its results are supported by extensive preclinical data. Detection on parenchymal A β plaques following a single systemic administration confirmed that aducanumab penetrates the brain to a sufficient extent to allow accumulation on A β plaques. This is consistent with earlier findings showing that, in the presence of significant A β deposition, plaque-binding antibodies can be detected bound to the target over an extended period^{14,21}. The minimal effective dose upon repeated systemic administration of ^{ch}aducanumab in transgenic mice was 3 mg kg⁻¹ (corresponding to minimally effective concentrations of 13.8 \pm 1.9 μ g ml⁻¹ in plasma and 99.8 \pm 30.0 ng g⁻¹ in brain) with reductions of A β ₄₂ in soluble and insoluble brain fractions of approximately 50%, and reductions in A β plaque of approximately 40%. Since exposure at 3 mg kg⁻¹ in animals and humans is approximately equivalent, the observed dose-response in the model was consistent with the clinical doses that led to reductions in amyloid PET SUVR. ^{ch}aducanumab cleared plaques of all sizes, suggesting that aducanumab triggered clearance of pre-existing A β plaques and prevented formation of new plaques.

In transgenic mice, aducanumab preferentially bound to parenchymal A β over vascular A β deposits, consistent with the lack of effect on vascular A β following chronic dosing. The effect of anti-A β antibody therapies on the vascular A β compartment could be related to micro-haemorrhages or oedema in transgenic mice, and may relate to ARIA in clinical trials²². Nevertheless, the preferential binding of aducanumab to parenchymal versus vascular A β may have been critical in allowing the use of relatively high doses in the clinical study so as to achieve robust target engagement in the brains of patients with AD.

Several mechanisms may be involved in aducanumab's A β -lowering activity. The clearance of A β deposits was accompanied by enhanced recruitment of microglia. Together with the reduced potency of the aglycosylated form of ^{ch}aducanumab (data not shown), and the *ex vivo* phagocytosis data, this suggests that Fc γ R-mediated microglial recruitment and phagocytosis played an important role in A β clearance in these models. Activated microglia appeared to encapsulate the remaining central dense core of plaques in treated animals, possibly isolating them from the surrounding neuropil. It is commonly thought that soluble A β oligomers, rather than monomers or plaques, may be the primary toxic species^{23,24}. Considering that A β plaques might be a source of A β oligomers^{25–28}, this suggests that treatment with aducanumab might slow their release into the neuropil, thereby limiting their toxic effect on neurons²⁹. In fact, chronic dosing of 18-month-old Tg2576 transgenic mice with ^{ch}aducanumab led to normalization of neuritic calcium overload in the brain³⁰. Other studies have linked calcium dyshomeostasis in neurons and microglia to binding of A β oligomers to metabotropic receptors^{31–33}. Aducanumab binding to soluble A β oligomers may prevent their interaction with those receptors, thereby preventing the detrimental effect of membrane depolarization. Restoration of this functional endpoint suggests that aducanumab treatment may lead to beneficial effects on neuronal network function underlying cognitive deficits.

Together, the clinical and preclinical data support continued development of aducanumab as a disease-modifying treatment for AD. The clinical study results provide robust support to the biological hypothesis that treatment with aducanumab reduces brain A β plaques and, more importantly, to the clinical hypothesis that A β plaque reduction confers clinical benefit. This concurs with preclinical data demonstrating brain penetration, target engagement, and dose-dependent clearance of A β plaques in transgenic mice. The clinical effects of aducanumab need to be confirmed in larger studies. Both the long-term extension (LTE) phase of this study and phase 3 development are ongoing.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 July 2015; accepted 21 July 2016.

Published online 24 August 2016.

- Hardy, J. & Selkoe, D. J. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science* **297**, 353–356 (2002).
- Hardy, J. A. & Higgins, G. A. Alzheimer's disease: the amyloid cascade hypothesis. *Science* **256**, 184–185 (1992).
- Ising, C., Stanley, M. & Holtzman, D. M. Current thinking on the mechanistic basis of Alzheimer's and implications for drug development. *Clin. Pharmacol. Ther.* **98**, 469–471 (2015).
- Selkoe, D. J. The therapeutics of Alzheimer's disease: where we stand and where we are heading. *Ann. Neurol.* **74**, 328–336 (2013).
- Cummings, J. L., Morstorf, T. & Zhong, K. Alzheimer's disease drug-development pipeline: few candidates, frequent failures. *Alzheimers Res. Ther.* **6**, 37 (2014).
- Doody, R. S. *et al.* Phase 3 trials of solanezumab for mild-to-moderate Alzheimer's disease. *N. Engl. J. Med.* **370**, 311–321 (2014).
- Salloway, S. *et al.* Two phase 3 trials of bapineuzumab in mild-to-moderate Alzheimer's disease. *N. Engl. J. Med.* **370**, 322–333 (2014).
- Ferrero, J. *et al.* First-in-human, double-blind, placebo-controlled, single-dose escalation study of aducanumab (BIB037) in mild-to-moderate Alzheimer's disease. *Alzheimers Dement. (N Y)* (in press).
- Sevigny, J. *et al.* Amyloid PET screening for enrichment of early-stage Alzheimer disease clinical trials: experience in a phase 1b clinical trial. *Alzheimer Dis. Assoc. Disord.* **30**, 1–7 (2016).
- Landau, S. M. *et al.* Amyloid- β imaging with Pittsburgh compound B and florbetapir: comparing radiotracers and quantification methods. *J. Nucl. Med.* **54**, 70–77 (2013).
- Banks, W. A. *et al.* Passage of amyloid beta protein antibody across the blood-brain barrier in a mouse model of Alzheimer's disease. *Peptides* **23**, 2223–2226 (2002).
- Levites, Y. *et al.* Insights into the mechanisms of action of anti-A β antibodies in Alzheimer's disease mouse models. *FASEB J.* **20**, 2576–2578 (2006).
- Bard, F. *et al.* Peripherally administered antibodies against amyloid beta-peptide enter the central nervous system and reduce pathology in a mouse model of Alzheimer disease. *Nat. Med.* **6**, 916–919 (2000).
- Bohrmann, B. *et al.* Gantenerumab: a novel human anti-A β antibody demonstrates sustained cerebral amyloid- β binding and elicits cell-mediated removal of human amyloid- β . *J. Alzheimers Dis.* **28**, 49–69 (2012).
- Villemagne, V. L. *et al.* Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study. *Lancet Neurol.* **12**, 357–367 (2013).
- Ostrowitzki, S. *et al.* Mechanism of amyloid removal in patients with Alzheimer disease treated with gantenerumab. *Arch. Neurol.* **69**, 198–207 (2012).
- Sperling, R. *et al.* Amyloid-related imaging abnormalities in patients with Alzheimer's disease treated with bapineuzumab: a retrospective analysis. *Lancet Neurol.* **11**, 241–249 (2012).
- Sperling, R. A. *et al.* Amyloid-related imaging abnormalities in amyloid-modifying therapeutic trials: recommendations from the Alzheimer's Association Research Roundtable Workgroup. *Alzheimers Dement.* **7**, 367–385 (2011).
- Barakos, J. *et al.* MR imaging features of amyloid-related imaging abnormalities. *AJNR Am. J. Neuroradiol.* **34**, 1958–1965 (2013).
- Zago, W. *et al.* Vascular alterations in PDAPP mice after anti-A β immunotherapy: Implications for amyloid-related imaging abnormalities. *Alzheimers Dement.* **9** (Suppl), S105–S115 (2013).
- Wang, A., Das, P., Switzer, R. C., III, Golde, T. E. & Jankowsky, J. L. Robust amyloid clearance in a mouse model of Alzheimer's disease provides novel insights into the mechanism of amyloid-beta immunotherapy. *J. Neurosci.* **31**, 4124–4136 (2011).
- Boche, D. *et al.* Consequence of A β immunization on the vasculature of human Alzheimer's disease brain. *Brain* **131**, 3299–3310 (2008).
- Haass, C. & Selkoe, D. J. Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid beta-peptide. *Nat. Rev. Mol. Cell Biol.* **8**, 101–112 (2007).
- Kayed, R. & Lasagna-Reeves, C. A. Molecular mechanisms of amyloid oligomers toxicity. *J. Alzheimers Dis.* **33** (Suppl 1), S67–S78 (2013).
- Benilova, I., Karran, E. & De Strooper, B. The toxic A β oligomer and Alzheimer's disease: an emperor in need of clothes. *Nat. Neurosci.* **15**, 349–357 (2012).
- Koffie, R. M. *et al.* Oligomeric amyloid beta associates with postsynaptic densities and correlates with excitatory synapse loss near senile plaques. *Proc. Natl Acad. Sci. USA* **106**, 4012–4017 (2009).
- Shankar, G. M. *et al.* Amyloid-beta protein dimers isolated directly from Alzheimer's brains impair synaptic plasticity and memory. *Nat. Med.* **14**, 837–842 (2008).
- Condello, C., Yuan, P., Schain, A. & Grutzendler, J. Microglia constitute a barrier that prevents neurotoxic protofibrillar A β 42 hotspots around plaques. *Nat. Commun.* **6**, 6176 (2015).
- Jin, M. *et al.* Soluble amyloid beta-protein dimers isolated from Alzheimer cortex directly induce tau hyperphosphorylation and neuritic degeneration. *Proc. Natl Acad. Sci. USA* **108**, 5819–5824 (2011).
- Kastanenka, K. *et al.* Amelioration of calcium dyshomeostasis by immunotherapy with BIB037 in Tg2576 mice. *Alzheimers Dement.* **9**, P508 (2013).

31. Jarosz-Griffiths, H. H., Noble, E., Rushworth, J. V. & Hooper, N. M. Amyloid- β receptors: the good, the bad, and the prion protein. *J. Biol. Chem.* **291**, 3174–3183 (2016).
32. Morkuniene, R. *et al.* Small A β 1-42 oligomer-induced membrane depolarization of neuronal and microglial cells: role of *N*-methyl-D-aspartate receptors. *J. Neurosci. Res.* **93**, 475–486 (2015).
33. Um, J. W. *et al.* Metabotropic glutamate receptor 5 is a coreceptor for Alzheimer $\alpha\beta$ oligomer bound to cellular prion protein. *Neuron* **79**, 887–902 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements These studies were funded by Biogen. The authors thank the patients and their family members participating in the aducanumab studies, and the PRIME investigators (Supplementary Information) and staff conducting these studies. Medical writing support, under direction of the authors, was provided by A. Smith at Complete Medical Communications, and was funded by Biogen. We thank N. Pederson, J. Dolnikova and E. Garber for help in generating the recombinant antibodies, D. Fahrner, C. Quigley, M. Themeles, X. Zhang and P. Auluck for help in generating the histological data, and K. Mack for editorial support and coordination of the authors in combining the preclinical and clinical work in this manuscript.

Author Contributions T.B., P.H.W., M.M., T.E., K.R., J.G. and R.M.N. designed the preclinical studies, and J.S., Y.L., J.G., J.F., C.H., R.M.N. and A.S. designed the clinical study. P.C. led the imaging implementation for the clinical study. T.C. and J.O. were clinical study statisticians. T.B., P.H.W., M.M., R.D., F.Q., M.A., M.L., S.C., M.S.B., O.Q.-M., R.H.S., H.M.A., T.E., J.G. and R.M.N. generated, analysed, and/or interpreted data from preclinical studies. T.B., P.H.W., M.M., R.D., F.Q., M.A., M.L., S.C., M.S.B., O.Q.-M., R.H.S., H.M.A., T.E., K.R., J.G., C.H., R.M.N. and A.S. critically reviewed preclinical sections of the manuscript. J.S., P.C., L.W., S.S., T.C., Y.L., J.O., J.F., Y.H., A.M., J.G., C.H., R.M.N. and A.S. analysed and interpreted clinical study data and critically reviewed clinical sections of the manuscript. All authors approved the final version of the manuscript for submission. Biogen and Neurimmune reviewed and provided feedback on the paper. The authors had full editorial control of the paper, and provided their final approval of all content.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.S. (alfred.sandrock@biogen.com).

Reviewer Information *Nature* thanks L. Lannfelt, R. Thomas and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Clinical study subjects. Patients were screened for inclusion in three stages. First, patients were evaluated on demographic, and clinical and laboratory criteria, including being between 50–90 years of age, and meeting clinical criteria for either prodromal or mild AD, as determined by the investigator. The criteria for prodromal AD were: MMSE score between 24–30 (inclusive), a spontaneous memory complaint, objective memory loss defined as a free recall score of ≤ 27 on the FCSRT³⁴, a global CDR score of 0.5, absence of significant levels of impairment in other cognitive domains and essentially preserved activities of daily living, and an absence of dementia³⁵. The criteria for mild AD were: MMSE score between 20–26 (inclusive), a global CDR of 0.5 or 1.0, and meeting the National Institute on Aging–Alzheimer's Association core clinical criteria for probable AD³⁶. Second, patients who remained eligible underwent MRI to exclude those with confounding pathology, including acute or sub-acute micro- or macro-haemorrhage, prior macro-haemorrhage, >4 micro-haemorrhages, superficial siderosis or any finding that might be a contributing cause of the patient's dementia, pose a risk to the patient, or prevent a satisfactory MRI assessment for safety monitoring. Third, remaining eligible patients underwent a florbetapir PET scan, and those with a positive scan based on a visual assessment, as determined by a qualified reader, were eligible. The A β PET screening process has been described in a separate publication⁹. Stable use of most concomitant background medications was permitted and, in the case of cholinesterase inhibitors and/or memantine, patients were required to be on a stable dose for a minimum of 4 weeks before screening with no adjustment of dosing during the double-blind phase of the study. Patients were excluded if they had a medical condition that might be a contributing cause of cognitive impairment.

Clinical study design. This was a multicentre, randomized, 12-month, double-blind, placebo-controlled, multiple-dose study of aducanumab followed by a 42-month, dose-blinded LTE study in patients with either prodromal or mild AD who were A β PET-positive (ClinicalTrials.gov identifier NCT01677572). The primary objective was to evaluate the safety and tolerability of multiple doses of aducanumab in patients with prodromal AD or mild AD dementia. The secondary objectives were to: (i) assess the effect on cerebral A β plaque content as measured by ¹⁸F-florbetapir PET imaging at week 26; (ii) assess the multiple-dose serum concentrations of aducanumab; and (iii) evaluate the immunogenicity of aducanumab after multiple-dose administration. The key exploratory objectives were assessments of the effect of aducanumab on the following: the clinical progression of AD as measured by change from baseline on the CDR-SB, a NTB, and the FCSRT; disease-related biomarkers in blood, cerebral A β plaque content as measured by ¹⁸F-florbetapir PET imaging at week 54; and cerebral A β plaque content by ApoE $\epsilon 4$ carrier status (carrier/non-carrier). Other exploratory endpoints were change from baseline on the Neuropsychiatric Inventory Questionnaire, Cognitive Drug Research computerized test battery, volumetric MRI, and, in a subset of patients, glucose metabolism as measured by fluorodeoxyglucose PET, functional connectivity by task-free functional MRI, cerebral blood flow by arterial spin labelling MRI, and disease-related biomarkers in cerebrospinal fluid. MMSE was another exploratory assessment.

During the 12-month, double-blind, placebo-controlled phase, patients received aducanumab or placebo by IV infusion once every 4 weeks for 52 weeks. In a staggered, parallel-group design, the treatment arms were enrolled as follows: first Arms 1–3 (aducanumab 1 mg kg⁻¹ ($n=30$); aducanumab 3 mg kg⁻¹ ($n=30$); placebo ($n=20$), respectively) in parallel. Once enrolment was open, Arms 4 and 5 (aducanumab up to 10 mg kg⁻¹ ($n=30$) (actual dose 10 mg kg⁻¹); placebo ($n=10$), respectively) were enrolled in parallel with Arms 1–3. Once enrolment in Arms 1–5 was complete, enrolment in Arms 6 and 7 (aducanumab up to 30 mg kg⁻¹ ($n=30$) (actual dose 6 mg kg⁻¹); placebo ($n=10$), respectively) began. The trial was initially designed to dose up to 30 mg kg⁻¹, but when ARIA were detected at 10 mg kg⁻¹ it was decided not to proceed to doses higher than 10 mg kg⁻¹ with repeated infusions. Dose escalation in Arms 4 and 5, and then Arms 6 and 7, was based on review of existing safety, tolerability, and pharmacokinetic data, and recommendation of the external Data Monitoring Committee. Patients were randomized (using a centralized interactive voice and web response System (IXRS)) to a treatment group within Arms 1–3, 4 and 5, or 6 and 7, stratified by ApoE $\epsilon 4$ status (carrier or non-carrier). Patient enrolment was monitored so that the ratio of ApoE $\epsilon 4$ carriers to non-carriers was no more than 2:1 and no less than 1:2. During the overlap in enrolment of Arms 1–3 and Arms 4 and 5, patients were randomized using a minimization algorithm. Patients who discontinued study treatment for any reason were encouraged to remain in the study and complete all assessments during the double-blind period. Patients completing the double-blind period and meeting certain eligibility criteria entered the LTE. After enrolment on Arms 6 and 7 were completed, the protocol was amended to include a titration arm and a corresponding placebo group—Arms 8 and 9. Both the LTE and Arms 8 and 9 are ongoing and were not part of this interim analysis.

Investigators, study site staff (except for a designated pharmacist/technician), and study patients were blinded to the patients' randomized treatment assignment

for the placebo-controlled period. Only the designated pharmacist/technician at each site was aware of the assigned treatment for each patient. Aducanumab was supplied as a sterile clear-to-yellow solution for IV infusion at a dose of 200 mg in 4 ml. For patients randomized to receive aducanumab, undiluted aducanumab (required volume based on patient weight) was added to a 100 ml 0.9% saline bag to reach the assigned dose (an equivalent amount of saline was first withdrawn so that the final total volume of all IV bags was identical). All IV bags (active and placebo (100 ml 0.9% saline)) were covered with a sealed brown light-protective bag to maintain blinding with a label including protocol and patient randomization number.

Cases of ARIA were managed in accordance with protocol-defined rules using centrally read MRI findings coupled with clinical symptoms, if present. The rules were consistent with the guidelines published by the Alzheimer Association Research Roundtable Working Group¹⁸. Briefly, patients developing mild ARIA-E or ARIA-H (≤ 4 incident micro-haemorrhages) without clinical symptoms could continue at the same dose; patients developing moderate or severe ARIA-E without clinical symptoms, or those with ARIA-E accompanied by mild clinical symptoms, could suspend treatment and resume at the next lower dose level once ARIA (and symptoms, if any) resolved. Patients who developed ARIA-E or ARIA-H (≤ 4 incident micro-haemorrhages) accompanied by moderate, severe, or serious clinical symptoms, >4 incident micro-haemorrhages, any incident macro-haemorrhage, or >1 incident haemosiderosis at any time during the study were to permanently discontinue treatment.

The study was conducted in accordance with the Declaration of Helsinki, and the International Conference on Harmonisation and Good Clinical Practice guidelines, and had ethics committee approval at each participating site. All patients provided written informed consent.

Clinical study assessments. Amyloid plaque content, as measured by florbetapir PET imaging, was assessed at screening, and at weeks 26 and 54. Detailed PET scanning protocols have been described in a separate publication⁹. Briefly, for each florbetapir scan, a dose of 370 MBq was injected intravenously, with PET scanning starting around 50 min later and continuing for approximately 20 min.

Visual reads, the basis for meeting the inclusion criterion of a positive A β PET scan, were based upon PET image data, with the registered MRI and fused PET/MRI data providing supplementary anatomical information. Scans were independently interpreted by two board-certified neuroradiologists who, in accordance with the Amyvid Prescribing Information³⁷, had successfully completed a training programme (provided by the manufacturer using either an in-person tutorial or an electronic process). Images were designated as positive or negative, following guidelines described in the Amyvid Prescribing Information³⁷.

A composite cortical SUVR was computed using a volume-weighted average across six brain regions of interest (frontal, parietal, lateral temporal and sensorimotor, anterior, and posterior cingulate cortices), as previously described¹⁶, normalized to whole cerebellar activity^{10,38}.

Clinical tests including the CDR and an NTB (comprising RAVLT Immediate and Delayed Recall, Wechsler Memory Scale Verbal Pair Associate Learning Test Immediate and Delayed Recall, Delis–Kaplan Executive Function System Verbal Fluency Conditions 1 and 2, and the Wechsler Adult Intelligence Scale Fourth Edition Symbol Search and Coding Subsets) were performed during screening and at weeks 26 and 54. The FCSRT was performed at screening and at week 52. These clinical tests were administered by a trained, certified clinician or rater experienced in the assessment of patients with cognitive deficits. When possible, the same rater would administer a given test across all visits. In order to maintain blinding to adverse events, raters were not permitted to perform other clinical assessments, and were blinded to other clinical and safety assessments. The rater who conducted the CDR for a patient could not complete any other rating scales for that same patient, and was blinded to the results of all other cognitive scales.

The following safety assessments were performed at regular intervals: physical examination, neurological examination, vital signs, electrocardiogram, and laboratory safety assessments. During the placebo-controlled period, brain MRI was performed at screening and at weeks 6, 18, 30, 42, and 54, and end of study or termination. The MMSE was completed at screening, and at weeks 24, 52, and end of study or termination, and, in patients who developed ARIA, at every scheduled visit until ARIA resolved.

The concentrations of aducanumab in serum and presence of anti-aducanumab antibodies were determined using validated ELISA techniques (Supplementary Information).

Statistical analysis in the clinical study. This interim analysis included all patients randomized to a fixed-dose regimen and completing the double-blind period of the study (data cut-off February 2015). For all analyses, all patients assigned to placebo were treated as a single group. The safety population was defined as all patients who were randomized and received at least one dose of study treatment. Adverse events were coded using the Medical Dictionary for Regulatory Activities

classification. The pharmacodynamic and pharmacokinetic populations were defined as all patients who were randomized, received at least one dose of study treatment, and had at least one post-baseline assessment of the pharmacodynamic parameter or at least one measurable aducanumab concentration in serum, respectively.

The primary analysis of the pharmacodynamic and efficacy data was based on Analysis of Covariance (ANCOVA), adjusting for baseline and ApoE ϵ 4 status (carrier and non-carrier) using observed data. No imputation was performed for missing data. For each time point, adjusted means for each treatment, pairwise adjusted differences with placebo, 95% confidence intervals for the pairwise differences, and associated nominal *P* values for comparison were calculated. No adjustments were made for multiple comparisons/multiple interim analyses. Dose-response was tested using a linear contrast from the ANCOVA model. The linear contrast test is sensitive to a variety of positive dose-response shapes, including a linear dose-response relationship. This served as the primary analysis for the cognition analyses. To account for missing data, a MMRM was used as a sensitivity analysis for the longitudinal data change from baseline data, adjusting for baseline and ApoE ϵ 4 status (carrier and non-carrier). Visit and treatment group were treated as categorical variables in the model along with their interactions. An unstructured covariance matrix was assumed to model the within-patient variability. This model imposes no assumptions on mean trend and correlation structure, and is considered robust.

Subgroup analyses were performed for change from baseline A β PET and change from baseline for cognition measures (CDR-SB and MMSE) for baseline clinical stage and ApoE ϵ 4 status (carrier and non-carrier). The subgroup analysis of the pharmacodynamic and efficacy data was based on ANCOVA, adjusting for baseline and ApoE ϵ 4 status (carrier and non-carrier) (for baseline clinical stage only) using observed data.

Serum pharmacokinetics were determined by nonlinear mixed effects model (NONMEM) approach. Sparse samples in the multiple-ascending-dose study and intensive samples from an earlier single-ascending-dose study⁸ were combined to construct a population pharmacokinetic model. The model was built in NONMEM software using the first-order conditional estimation with interaction method. Cumulative AUC up to month 12 was estimated for each patient. The plasma terminal elimination half-life was estimated in the pharmacokinetic analysis population. The analysis population for the primary immunogenicity analysis was defined as all patients who were randomized, received study treatment, and had at least one post-dose immunogenicity sample evaluated for immunogenicity.

Interim analyses were specified in the protocol for the purpose of planning future studies; no changes were to be made for this study based on the interim analysis results.

A sample size of 30 patients per treatment group would provide more than 90% power to detect a treatment difference of 1 standard deviation with respect to the reduction of A β from baseline, based on comparison of each aducanumab group with placebo, at a two-sided significance level of 0.05, and assuming a dropout rate of 20%.

Transgenic mouse studies. Penetration of aducanumab into the brain and target engagement were assessed in 22-month-old female Tg2576 mice following a single dose of aducanumab at 30 mg kg⁻¹ administered i.p. ('Target engagement study'; *n* = 4–5 per time point). The ability of aducanumab to reduce A β burden was assessed following chronic treatment of 9-month-old male and female Tg2576 transgenic mice dosed weekly i.p. for 6 months with PBS or 0.3, 1, 3, 10, or 30 mg kg⁻¹ of the murine chimaeric variant ^{ch}aducanumab ('Chronic efficacy study'; *n* = 20–55 per treatment group). An additional dosing study ('Chronic efficacy study with Agly'; *n* = 12–14 per treatment group) comparing the plaque clearing ability of ^{ch}aducanumab to that of an effector function-impaired variant (^{ch}aducanumab-Agly) was conducted using a similar study design (chronic treatment of 9.5-month-old Tg2576 transgenic mice dosed weekly i.p. for 6 months with PBS or 3 mg kg⁻¹ of ^{ch}aducanumab or ^{ch}aducanumab-Agly).

Mice were killed following anaesthesia with ketamine/xylazine (100/10 mg kg⁻¹ i.p.). Blood was collected by cardiac puncture, and mice were perfused with ice-cold heparinized saline (0.9%) using a peristaltic pump. The brain was removed and halved along the medio-sagittal line. The right hemisphere was frozen on dry ice and stored at -80°C for biochemical analysis. The left hemisphere was fixed by immersion in 10% neutral buffered formalin.

Size of the treatment groups was determined to take into account natural mortality (10–20%) and high inter-animal variability specific to the Tg2576 strain of mice. No animals were excluded from the analyses, unless the animal died prematurely. 'n' reported in the manuscript represents the number of animals in each group that were euthanized as scheduled at the end of the study. The allocation of animals to treatment groups took into account date of birth, gender, and weight at baseline. Each treatment group was balanced for mean age, gender, and mean weight. Dosing solutions were coded with letters so that all experimenters were

blinded to the treatment. The labelling of the samples collected did not reflect treatment group, so that experimenters processing and analysing the samples were still blinded. Codes were broken once all analyses were completed, including statistical analysis.

All in-life procedures were conducted in strict accordance with protocols approved by Biogen's Institutional Animal Care and Use Committee.

Biochemical measurements. Please see Supplementary Information.

Histological assessment. Please see Supplementary Information.

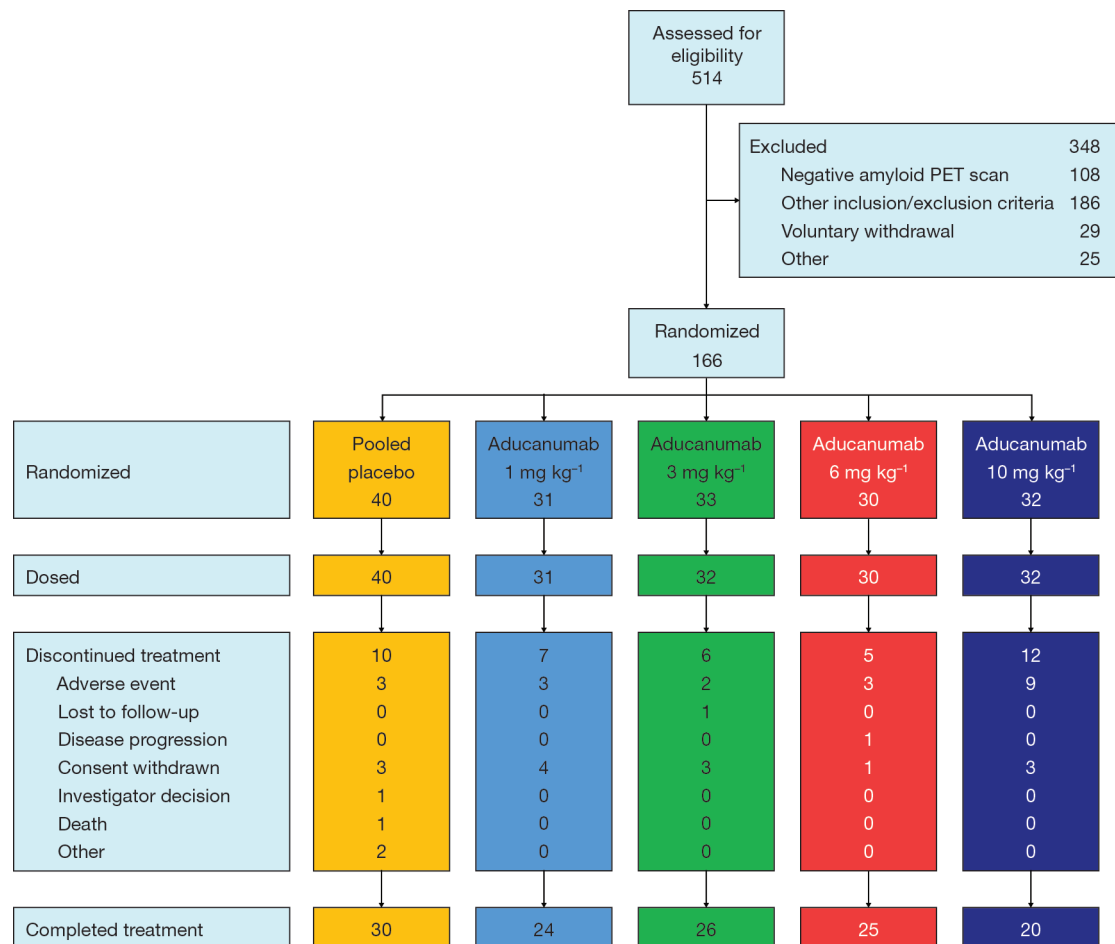
Preparation of different A β peptide conformations. Synthetic A β 1–42 (A β ₄₂) peptide (AnaSpec, Fremont, California, USA) was reconstituted in hexafluoro-isopropanol at a concentration of 1 mg/ml, aliquoted, air-dried, and vacuum-concentrated to form a film, and dissolved in dimethyl sulfoxide (DMSO) at a concentration of 5 mg/ml. A β ₄₂ oligomers and A β ₄₂ fibrils were prepared by diluting DMSO-reconstituted monomeric into PBS at a concentration of 100 μ g/ml and incubating at 37°C for at least 3 days and 1 week, respectively. The solution was centrifuged at 14,000g for 15 min at 4°C, and oligomers were recovered from the supernatant following the shorter incubation, whereas fibrils were recovered from the pellet following the longer incubation. For details on the biophysical characterization of high molecular weight A β ₄₂ aggregates, please see Supplementary Information.

In immunoprecipitation experiments, samples of freshly prepared monomeric, soluble oligomeric, or insoluble fibrillar A β ₄₂ were immunoprecipitated with ^{ch}aducanumab, 3D6 or a murine IgG2a control antibody (P1.17), dot-blotted onto a nitrocellulose membrane, and detected with biotinylated pan-A β antibody 6E10. Similar results were observed for ^{ch}aducanumab when immunoblotted with 3D6. **ELISA.** Please see Supplementary Information.

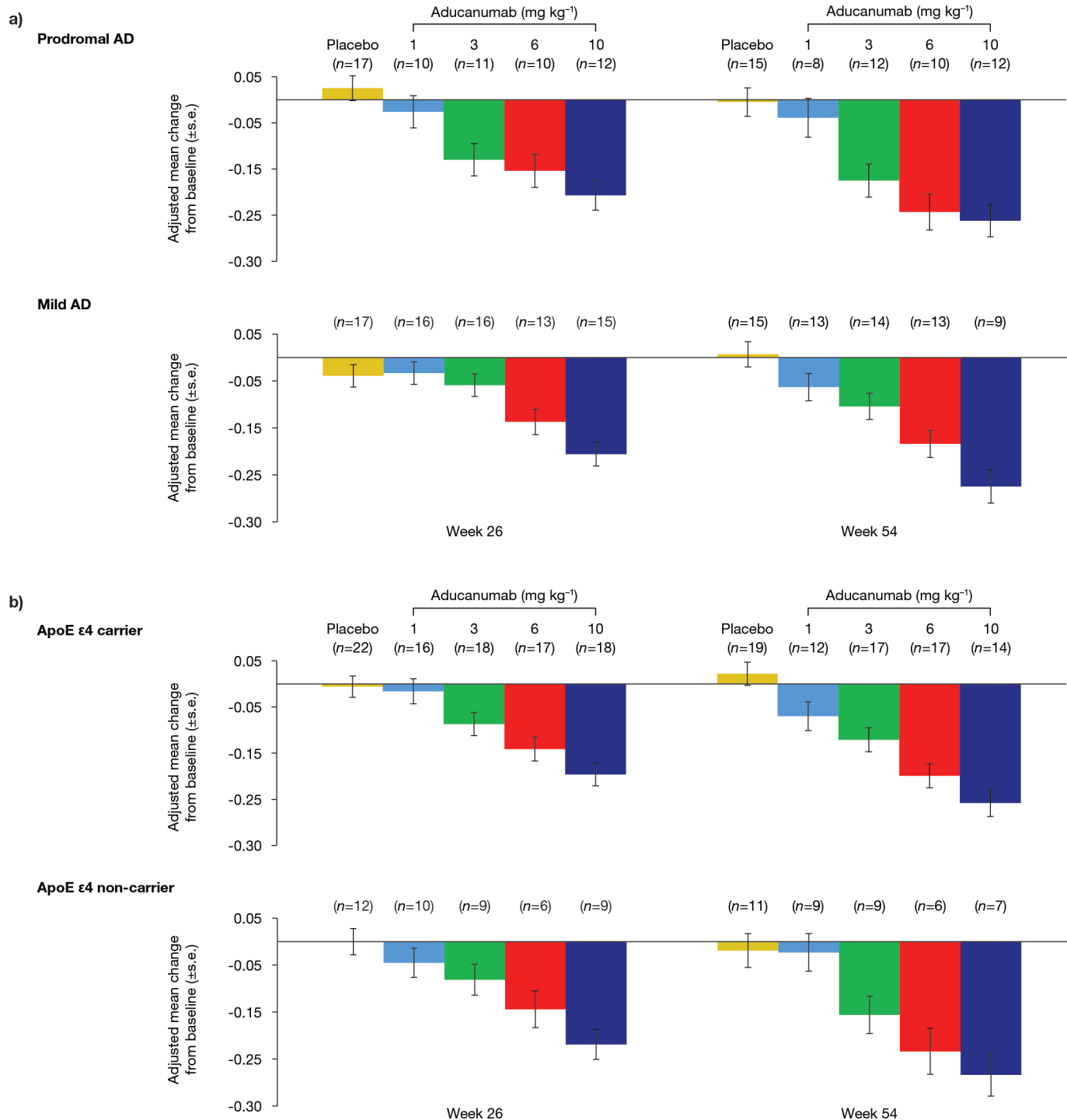
Antibody generation using reverse translational medicine. Aducanumab was derived from a de-identified blood lymphocyte library collected from healthy elderly subjects with no signs of cognitive impairment and cognitively impaired elderly subjects with unusually slow cognitive decline. Memory B cells, isolated from peripheral blood lymphocyte preparations by anti-CD22-mediated sorting were cultured on gamma-irradiated human peripheral blood mononuclear cell feeder layers. Supernatants from isolated B cells were screened for their ability to stain A β plaques on brain tissue sections, from either patients with AD or aged APP transgenic mice³⁹, and for their binding to aggregated forms of A β ₄₀ and A β ₄₂ *in vitro*. Positive hits meeting the above criteria were counter-screened to exclude clones cross-reacting with full-length APP expressed on stably transfected HEK293 cells (provided by U. Konietzko, University of Zurich, Switzerland; tested negative for mycoplasma contamination; not independently authenticated). Selected A β -reactive B-cell clones were subjected to cDNA cloning of IgG heavy and κ or λ light chain variable region sequences, and sub-cloned in expression constructs using Ig-framework specific primers for human variable heavy and light chain families in combination with human J-H segment-specific primers. Aducanumab was engineered to incorporate glycosylated human IgG1 heavy and human κ light chain constant domain sequences. A murine chimaeric IgG2a/ κ version of aducanumab (^{ch}aducanumab) was generated for use in chronic efficacy studies in APP transgenic mice. An aglycosylated variant of ^{ch}aducanumab (^{ch}aducanumab-Agly), incorporating a single point mutation (N297Q, using standard Kabat EU numbering) which eliminates N-glycosylation of the Fc region and severely reduces Fc γ R binding⁴⁰, was generated to test for Fc-related activities. The recombinant mouse IgG2b monoclonal antibody 3D6⁴¹ was used as a comparator in some studies.

Ex vivo phagocytosis assay. Please see Supplementary Information.

34. Derby, C. A. *et al.* Screening for predementia AD: time-dependent operating characteristics of episodic memory tests. *Neurology* **80**, 1307–1314 (2013).
35. Dubois, B. *et al.* Revising the definition of Alzheimer's disease: a new lexicon. *Lancet Neurol.* **9**, 1118–1127 (2010).
36. McKhann, G. M. *et al.* The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* **7**, 263–269 (2011).
37. Eli Lilly and Company. Amyvid Prescribing Information. <http://www.lilly.com>. (2013).
38. Clark, C. M. *et al.* Cerebral PET with florbetapir compared with neuropathology at autopsy for detection of neuritic amyloid- β plaques: a prospective cohort study. *Lancet Neurol.* **11**, 669–678 (2012).
39. Hock, C. *et al.* Generation of antibodies specific for beta-amyloid by vaccination of patients with Alzheimer disease. *Nat. Med.* **8**, 1270–1275 (2002).
40. Tao, M. H. & Morrison, S. L. Studies of aglycosylated chimeric mouse-human IgG. Role of carbohydrate in the structure and effector functions mediated by the human IgG constant region. *J. Immunol.* **143**, 2595–2601 (1989).
41. Johnson-Wood, K. *et al.* Amyloid precursor protein processing and A β deposition in a transgenic mouse model of Alzheimer disease. *Proc. Natl Acad. Sci. USA* **94**, 1550–1555 (1997).

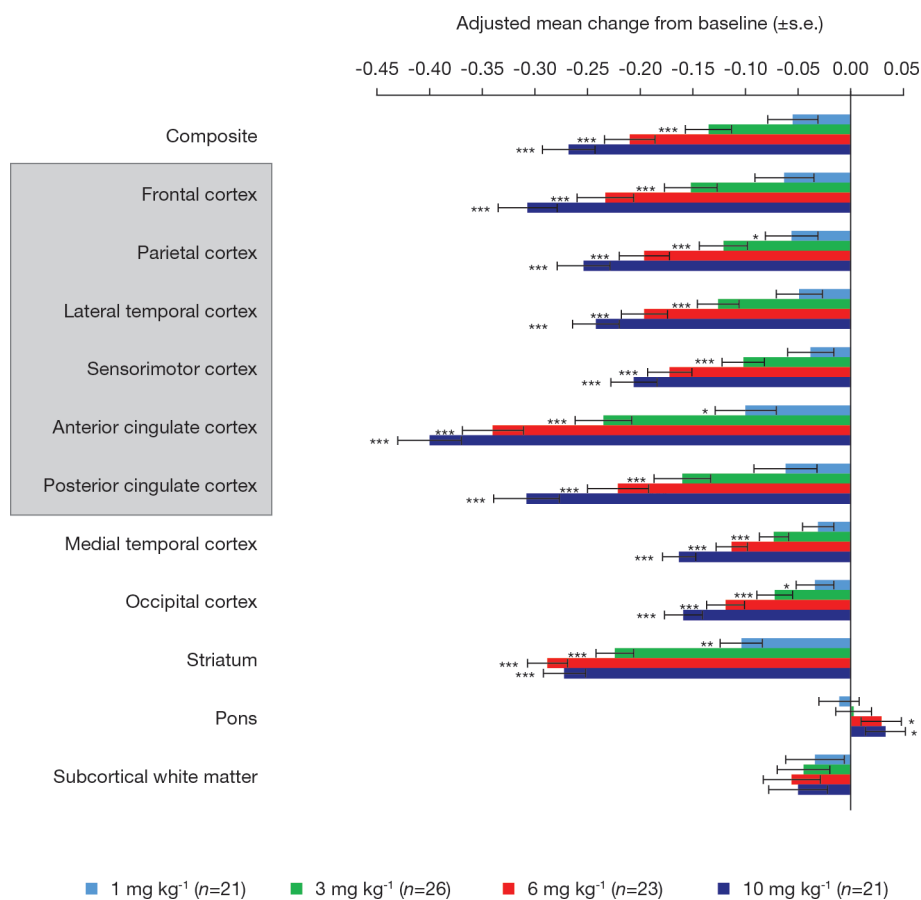


Extended Data Figure 1 | Participant accounting. PET, positron emission tomography.

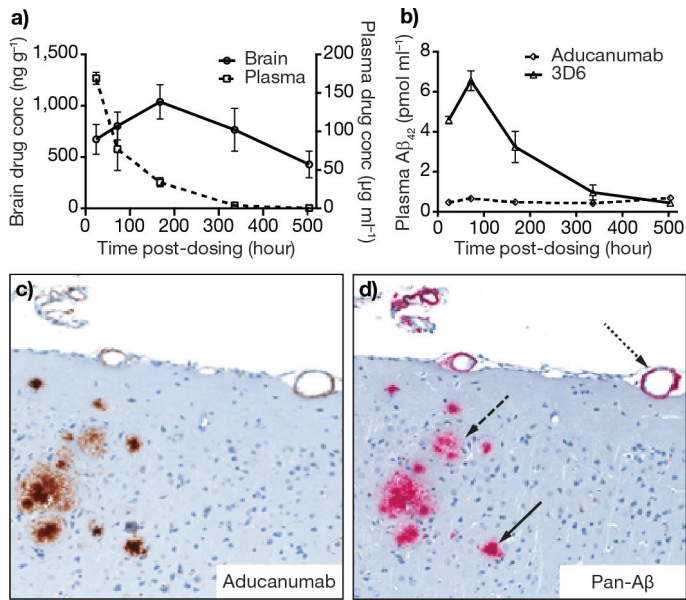


Extended Data Figure 2 | Amyloid plaque reduction with aducanumab by baseline clinical stage and baseline ApoE $\epsilon 4$ status. a, b, Analyses by baseline clinical stage were performed using ANCOVA for change from baseline with factors of: treatment, ApoE $\epsilon 4$ status (carrier and

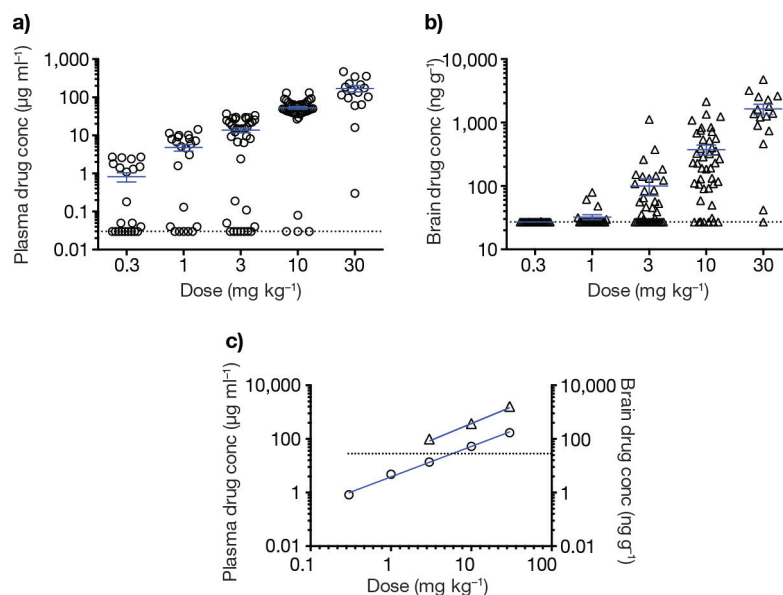
non-carrier) and baseline composite SUVR (a), and for analyses by ApoE $\epsilon 4$ status, using treatment and baseline composite SUVR (b). Adjusted mean \pm s.e. ApoE $\epsilon 4$, apolipoprotein E $\epsilon 4$ allele; SUVR, standard uptake value ratio.



Extended Data Figure 3 | Amyloid plaque reduction: regional analysis SUVR at week 54. The boxed area indicates the six regions included in the composite score. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$ versus placebo; two-sided tests with no adjustments for multiple comparisons. Adjusted mean \pm s.e. Analyses using ANCOVA. SUVR, standard uptake value ratio.



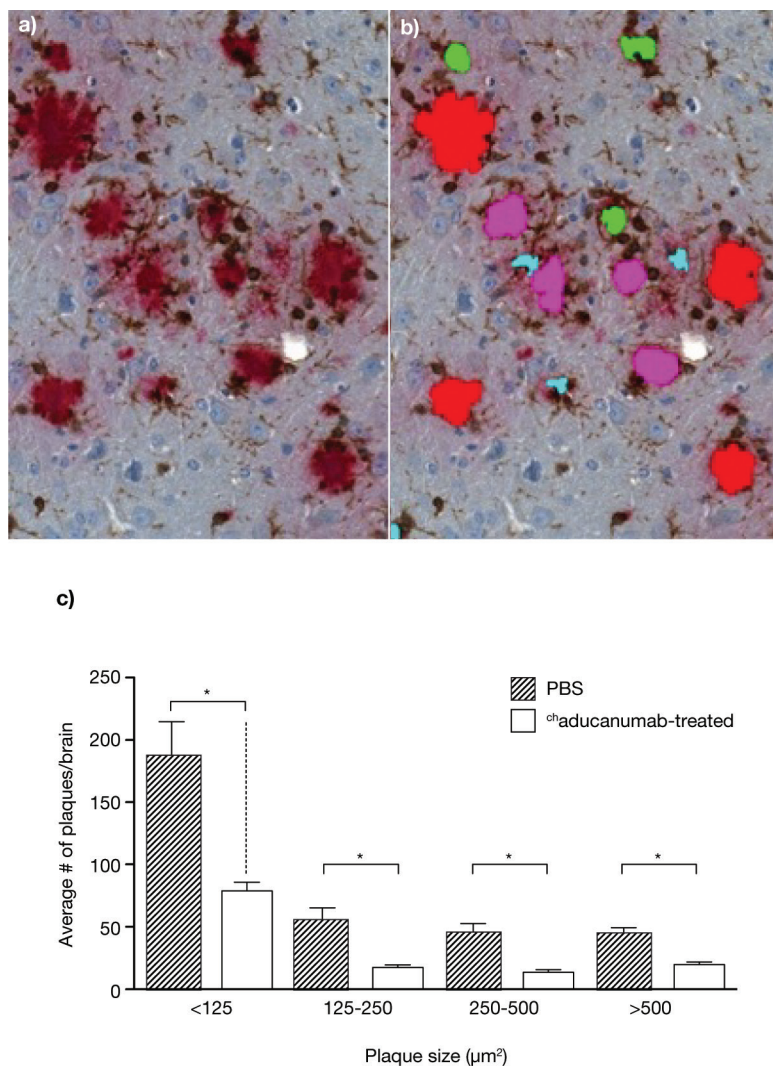
Extended Data Figure 4 | Brain penetration of aducanumab after a single intraperitoneal administration in 22-month-old Tg2576 transgenic mice. **a, b,** Aducanumab levels in plasma and brain (**a**), and plasma Aβ levels after a single dose (**b**; $n = 4-5$; mean \pm s.e.). **c, d,** *In vivo* binding of aducanumab to amyloid deposits detected using a human IgG-specific secondary antibody (**c**), and *ex vivo* immunostaining with a pan-Aβ antibody on consecutive section (**d**). Examples of a compact Aβ plaque (solid arrow), diffuse Aβ deposit (dashed arrow), and CAA lesion (dotted arrow). CAA, cerebral amyloid angiopathy.



Extended Data Figure 5 | Exposure following weekly dosing with ^{ch}aducanumab in 9.5- to 15.5-month-old Tg2576 transgenic mice.

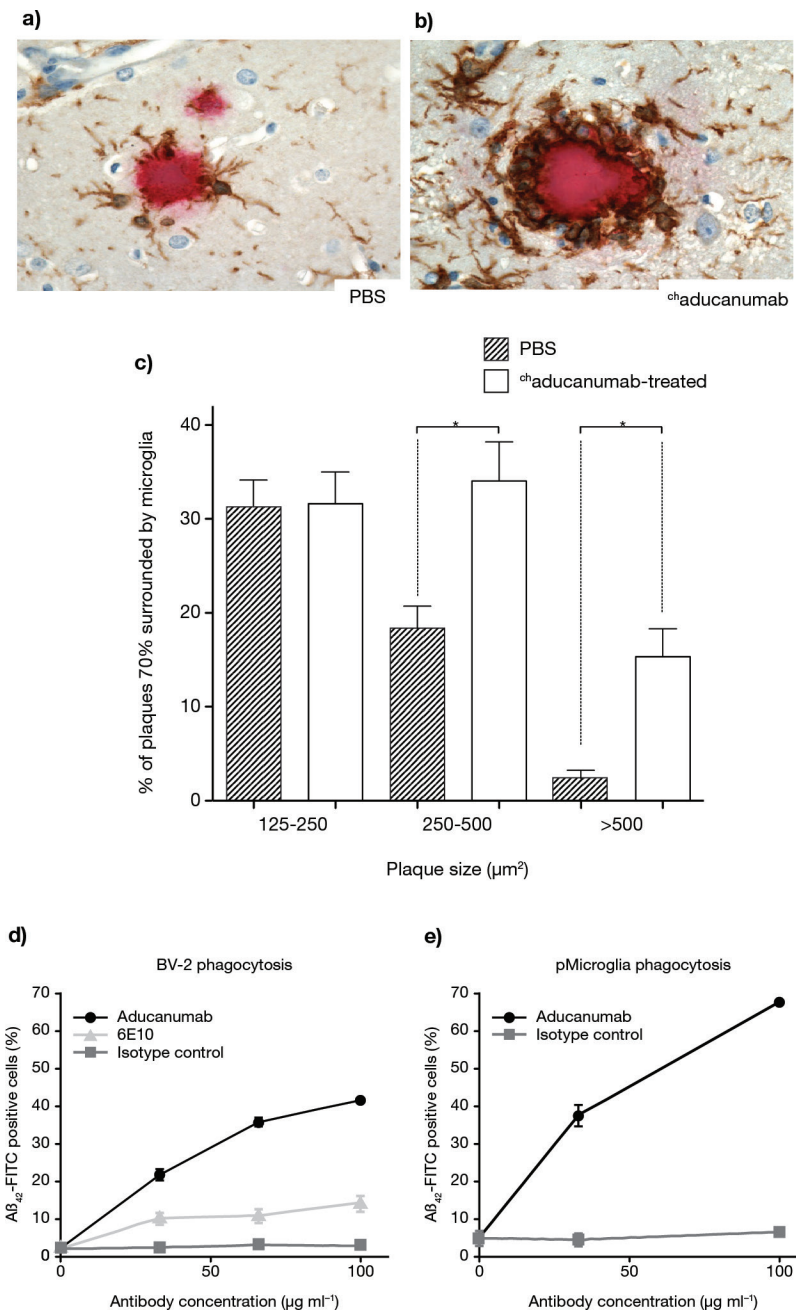
a, b, ^{ch}aducanumab concentrations in plasma (**a**), or DEA-soluble brain extract (**b**) were measured in samples collected 24 h after the last dose in the 'Chronic efficacy study'. Mean \pm s.e. Dotted lines represent the limits

of quantitation of each assay. **c,** Correlations of drug concentrations in plasma (open circles) or brain (open triangles) with administered dose. The average brain concentrations in the two groups receiving the lowest dose were below the limit of quantitation for that assay, which is indicated by a dotted line on the figure.



Extended Data Figure 6 | Treatment with chaducanumab affects plaques of all sizes. **a**, Following weekly dosing of chaducanumab in Tg2576 from 9.5–15.5 months of age, amyloid plaques were stained with 6E10 and quantified using Visiopharm software. **b**, Plaque size was defined by area, and coloured as follows: <125 μm^2 (cyan), 125–250 μm^2 (green), 250–500 μm^2 (pink), and >500 μm^2 (red). **c**, chaducanumab treatment was

associated with a significant decrease in plaque number in all size ranges relative to vehicle-treated controls, with reductions of 58%, 68%, 68%, and 53% in the number of plaques for the <125 μm^2 , 125–250 μm^2 , 250–500 μm^2 , and >500 μm^2 groups size, respectively. Mean \pm s.e.; statistically significant differences from vehicle for each size range are indicated with asterisks; $*P < 0.05$, Mann–Whitney test.



Extended Data Figure 7 | Enhanced recruitment of microglia to amyloid plaques following ^{ch}aducanumab treatment and engagement of Fc γ receptors. **a, b,** Brain sections from either PBS- or ^{ch}aducanumab-treated mice ('Chronic efficacy study'; 3 mg kg⁻¹ group) were immunostained for A β (6E10; red) and a marker of microglia (Iba1; brown). **c,** The area of individual amyloid plaques was measured, and Iba1-stained microglia were grouped into two categories, either associated with plaques (within 25 μm of a plaque) or not associated with plaques (>25 μm from a plaque). Plaques with circumferences \geq 70% surrounded by microglia were quantified and stratified based on plaque size. The fraction of plaques that were at least 70% surrounded

by microglia was significantly greater in the ^{ch}aducanumab-treated group (white bars) compared with the PBS control group (grey bars), for plaques \geq 250 μm^2 . Mean \pm s.e.; statistically significant differences from vehicle for each size range are indicated with asterisks; * P < 0.05, Bonferroni's post hoc test following one-way analysis of variance. All quantifications were done using the Visiopharm software. **d, e,** FITC-labelled A β_{42} fibrils were incubated with different concentrations of the antibodies before adding to BV-2 microglia cell line (**d**), or primary microglia (**e**) for phagocytosis experiment measuring uptake of A β_{42} fibrils into the cells by FACS analysis. Mean \pm s.d.

Extended Data Table 1 | Change from baseline in amyloid PET SUVR values (a secondary endpoint at 6 months), and in exploratory clinical endpoints at the end of the placebo-controlled period (6-month data also shown for amyloid PET)

Adjusted mean ± SE change from baseline for:	Aducanumab					p-value (dose-response)
	Placebo	1 mg kg ⁻¹	3 mg kg ⁻¹	6 mg kg ⁻¹	10 mg kg ⁻¹	
Amyloid PET SUVR values						
At 6 months	(n=34)	(n=26)	(n=27)	(n=23)	(n=27)	
	-0.005 ± 0.018	-0.030 ± 0.020	-0.087 ± 0.020**	-0.143 ± 0.022***	-0.205 ± 0.020***	<0.0001
At 1 year†	(n=30)	(n=21)	(n=26)	(n=23)	(n=21)	
	0.003 ± 0.021	-0.055 ± 0.024	-0.135 ± 0.022***	-0.210 ± 0.024***	-0.268 ± 0.025***	<0.0001
CDR-SB†	(n=31)	(n=23)	(n=27)	(n=26)	(n=23)	
	1.87 ± 0.41	1.72 ± 0.46	1.37 ± 0.43	1.11 ± 0.44	0.63 ± 0.47*	<0.05
MMSE‡	(n=32)	(n=25)	(n=26)	(n=26)	(n=25)	
	-2.81 ± 0.67	-2.18 ± 0.75	-0.70 ± 0.75*	-1.96 ± 0.75	-0.56 ± 0.76*	<0.05
NTB overall Z score†	(n=29)	(n=23)	(n=26)	(n=24)	(n=24)	
	-0.11 ± 0.08	-0.25 ± 0.09	-0.13 ± 0.08	-0.19 ± 0.09	-0.10 ± 0.09	NS
FCSRT: sum of free recall score‡	(n=31)	(n=23)	(n=25)	(n=25)	(n=25)	
	-2.33 ± 1.07	-1.63 ± 1.24	-1.25 ± 1.20	-4.04 ± 1.21	-0.69 ± 1.20	NS

* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$ versus placebo; two-sided tests with no adjustments for multiple comparisons.

[†]At week 54.

[‡]At week 52.

Analyses using ANCOVA. ApoE ϵ 4, apolipoprotein E ϵ 4 allele; CDR-SB, Clinical Dementia Rating—Sum of Boxes; FCSRT, Free and Cued Selective Reminding Test; MMSE, Mini-Mental State Examination; NS, not significant; NTB, neuropsychological test battery; SE, standard error; SUVR, standard uptake value ratio.

Extended Data Table 2 | Incidence of ARIA based on MRI data and ARIA-E patient disposition

	Placebo	Aducanumab			
		1 mg kg ⁻¹	3 mg kg ⁻¹	6 mg kg ⁻¹	10 mg kg ⁻¹
Number of dosed subjects with at least one post-baseline MRI	38	31	32	30	32
ApoE ε4 carrier	24	19	21	21	20
ApoE ε4 non-carrier	14	12	11	9	12
ARIA-E, n (%)	0	1 (3)	2 (6)	11 (37)	13 (41)
By ApoE ε4					
ApoE ε4 carrier	0	1 (5)	1 (5)	9 (43)	11 (55)
ApoE ε4 non-carrier	0	0	1 (9)	2 (22)	2 (17)
ARIA-E and:					
Continued treatment	0	0	2 (6)	8 (27)	5 (16)
Same dose	0	0	0	2 (7)	0
Dose reduced	0	0	2 (6)	6 (20)	5 (16)
Discontinued treatment	0	1 (3)	0	3 (10)	8 (25)
ApoE ε4 carrier	0	1 (5)	0	2 (10)	7 (35)
ApoE ε4 non-carrier	0	0	0	1 (11)	1 (8)
Isolated ARIA-H, n (%)	2 (5)	2 (6)	3 (9)	0	2 (6)
ApoE ε4 carrier	2 (8)	1 (5)	2 (10)	0	2 (10)
ApoE ε4 non-carrier	0	1 (8)	1 (9)	0	0
ARIA-E and ARIA-H, n (%)	0	1 (3)	1 (3)	5 (17)	8 (25)
ApoE ε4 carrier	0	1 (5)	1 (5)	5 (24)	7 (35)
ApoE ε4 non-carrier	0	0	0	0	1 (8)

ApoE ε4, apolipoprotein E ε4 allele; ARIA, amyloid-related imaging abnormalities; ARIA-E (oedema); ARIA-H (micro-haemorrhages, macro-haemorrhages, or superficial siderosis); MRI, magnetic resonance imaging.

Extended Data Table 3 | Pharmacokinetic data

	Aducanumab			
	1 mg kg ⁻¹	3 mg kg ⁻¹	6 mg kg ⁻¹	10 mg kg ⁻¹
PK analysis population (intent-to-treat)*	<i>n</i> =31	<i>n</i> =32	<i>n</i> =30	<i>n</i> =32
Cumulative AUC (µg.h/mL, mean ± SD)	47,078 ± 17,555	143,395 ± 59,986	251,535 ± 122,883	346,163 ± 198,603
Subjects who received all 14 planned doses	<i>n</i> =18	<i>n</i> =18/19†	<i>n</i> =16	<i>n</i> =14
C _{max ss} (µg/mL, mean ± SD)‡	21.2 ± 3.7	59.6 ± 19.6	123.8 ± 42.5	250.8 ± 33.5
Cumulative AUC (µg.h/mL, mean ± SD)	55,223 ± 11,529	169,457 ± 41,775	315,352 ± 76,300	524,511 ± 95,622

*Data include patients who missed doses.

†A total of 19 patients received all 14 doses but 1 patient missed the concentration measurement at Week 40 and so *n* = 18 for C_{max,ss} at 3 mg kg⁻¹ aducanumab.

‡The observed post-infusion concentrations at Week 40 were reported as steady-state C_{max}.

AUC, area under the concentration curve; C_{max,ss}, maximum concentration at steady state; PK, pharmacokinetic; SD, standard deviation.

Extended Data Table 4 | Change from baseline in amyloid PET SUVR values, CDR-SB, and MMSE at the end of the placebo-controlled period by absence/presence* of ARIA-E

Adjusted mean \pm SE for:	Treatment group (# without ARIA-E, # with ARIA-E)				
	Aducanumab				
	Placebo	1 mg kg ⁻¹	3 mg kg ⁻¹	6 mg kg ⁻¹	10 mg kg ⁻¹
Amyloid PET SUVR values [†]	(30, 0)	(21, 0)	(24, 2)	(17, 6)	(13, 8)
ARIA-E					
Absence	0.003 \pm 0.020	−0.056 \pm 0.024	−0.141 \pm 0.023	−0.243 \pm 0.027	−0.278 \pm 0.031
Presence	0.001 \pm 0.020	–	−0.069 \pm 0.075	−0.114 \pm 0.049	−0.263 \pm 0.040
CDR-SB [‡]	(31, 0)	(23, 0)	(25, 2)	(18, 8)	(14, 9)
ARIA-E					
Absence	1.84 \pm 0.42	1.72 \pm 0.48	1.33 \pm 0.47	1.11 \pm 0.54	0.78 \pm 0.61
Presence	1.95 \pm 0.35	–	2.04 \pm 1.38	1.18 \pm 0.73	0.67 \pm 0.67
MMSE [‡]	(32, 0)	(25, 0)	(24, 2)	(18, 8)	(16, 9)
ARIA-E					
Absence	−2.86 \pm 0.69	−2.20 \pm 0.77	−0.47 \pm 0.80	−1.82 \pm 0.91	−1.05 \pm 0.96
Presence	−2.60 \pm 0.69	–	−3.41 \pm 2.69	−1.95 \pm 1.42	0.83 \pm 1.35

*Since there were no ARIA-E events in the placebo group, the overall placebo group was used as the comparator in the subgroup analysis for presence of ARIA-E.

[†]At week 54.

[‡]At week 52.

Analyses based on observed data. Adjusted mean change and standard errors are based on an ANCOVA model for change from baseline with factors of treatment, laboratory ApoE ϵ 4 status (carrier and non-carrier), and baseline composite SUVR, CDR-SB, or MMSE, respectively. ARIA-E, amyloid-related imaging abnormalities (oedema); CDR-SB, Clinical Dementia Rating—Sum of Boxes; MMSE, Mini-Mental State Examination; PET, positron emission tomography; SE, standard error; SUVR, standard uptake value ratio.

A developmental coordinate of pluripotency among mice, monkeys and humans

Tomonori Nakamura^{1,2}, Ikuhiro Okamoto^{1,2}, Kotaro Sasaki^{1,2}, Yukihiro Yabuta^{1,2}, Chizuru Iwatani³, Hideaki Tsuchiya³, Yasunari Seita³, Shinichiro Nakamura³, Takuya Yamamoto^{4,5,6} & Mitinori Saitou^{1,2,4,5}

The epiblast (EPI) is the origin of all somatic and germ cells in mammals, and of pluripotent stem cells *in vitro*. To explore the ontogeny of human and primate pluripotency, here we perform comprehensive single-cell RNA sequencing for pre- and post-implantation EPI development in cynomolgus monkeys (*Macaca fascicularis*). We show that after specification in the blastocysts, EPI from cynomolgus monkeys (cyEPI) undergoes major transcriptome changes on implantation. Thereafter, while generating gastrulating cells, cyEPI stably maintains its transcriptome over a week, retains a unique set of pluripotency genes and acquires properties for ‘neuron differentiation’. Human and monkey pluripotent stem cells show the highest similarity to post-implantation late cyEPI, which, despite co-existing with gastrulating cells, bears characteristics of pre-gastrulating mouse EPI and epiblast-like cells *in vitro*. These findings not only reveal the divergence and coherence of EPI development, but also identify a developmental coordinate of the spectrum of pluripotency among key species, providing a basis for better regulation of human pluripotency *in vitro*.

The early embryonic development in mammals is decisive for the development of the pluripotent lineage, the EPI, which generates all of the somatic and the germ cell lineages^{1,2}. EPI is also a source of pluripotent stem cells (PSCs) *in vitro*, including embryonic stem cells (ESCs) derived from the pre-implantation EPI^{3,4} and epiblast stem cells derived from the post-implantation EPI^{5,6}, which, together with induced pluripotent stem cells (iPSCs)^{7,8}, are critical resources for a broad range of biomedical applications⁹.

Mammalian development has been studied almost exclusively using mice as a model organism. However, the mechanisms for mammalian development, including EPI development, are divergent among species¹⁰. Consequently, PSCs derived from EPI also have divergent properties; whereas mouse (m) ESCs have a naive pluripotency with unbiased potential for differentiation and chimaera contribution, human (h) or primate ESCs show similarity to mouse epiblast stem cells and exhibit a primed pluripotency with biased differentiation capacity and limited potential for chimaera contribution¹¹. For better understanding of hPSCs, investigation into the mechanism for human/primate embryonic development is critical. Such investigation, however, has been hampered owing to difficulties in analysing human and primate early post-implantation embryos. To gain insight into the mechanism of embryonic development in humans and primates and the properties of hPSCs, we performed a comprehensive analysis of the single-cell transcriptome during pre- and early post-implantation development in cynomolgus monkeys (*Macaca fascicularis*), a primate closely related to humans and suitable for biological experimentation (Extended Data Fig. 1a).

Pre- and post-implantation development

We isolated the metaphase-II-stage oocytes, fertilized them and cultured the resultant embryos¹² (Fig. 1a, Extended Data Fig. 1b–e). The blastocysts were observed from embryonic day (E) 5 (Fig. 1a, Extended Data Fig. 1d). We examined the expression of key markers¹⁰ by immunofluorescence analysis. Nearly all of the cells in blastocysts

at E6 were positive for OCT4, whereas NANOG was more restricted (Fig. 1b). From E7, OCT4 appeared to decline in outer cells, whereas NANOG was confined to a subset of OCT4⁺ inner cell mass (ICM) (Fig. 1b). At E9, both OCT4 and NANOG were confined to ICM cells (Fig. 1b, Extended Data Fig. 1f). Unlike in mice, but like in humans¹³, CDX2 was undetectable in blastocysts at E6, and became evident in outer cells at E7 (Extended Data Fig. 1f, g). Whereas GATA4 became detectable around ICM from E7 onwards (Fig. 1b, Extended Data Fig. 1f, g), GATA6 was widely expressed in blastocysts at E6 and in all cells except OCT4⁺ ICM from E7 onwards (Extended Data Fig. 1f, h). GATA4 and CDX2 showed mutually exclusive expression (Extended Data Fig. 1g), whereas GATA6 was expressed in virtually all of the CDX2⁺ cells (Fig. 1c). TFAP2C exhibited strong, weak, and no expression in GATA6⁺ outer cells, OCT4⁺GATA6[−] ICM, and GATA6⁺ ICM, respectively (Extended Data Fig. 1i). Thus, the ICM versus trophectoderm specification may manifest relatively late and the EPI versus hypoblast specification takes place at around E7 via a unique mechanism (Extended Data Fig. 1j).

We isolated early post-implantation embryos (E13, E14, E16, and E17) with ethical approval (Methods, Extended Data Fig. 2a, b). Consistent with previous literature on rhesus embryos^{14–17} (Extended Data Fig. 2c), implanted embryos exhibited disc-shaped, columnar EPI continuous with squamous amnionic cells (Fig. 2a, b, Extended Data Fig. 2d). Beneath EPI were a basement membrane and visceral endoderm, which was continuous with yolk-sac endoderm. EPI, amnion, and visceral/yolk-sac endoderm (VE/YE) were embedded in extra-embryonic mesenchyme (Fig. 2b, Extended Data Fig. 2d). Gastrulating cells above visceral endoderm were apparent in an E16 embryo (Fig. 2b). EPI was positive for OCT4 and NANOG, and VE/YE was positive for GATA4 and GATA6, and the gastrulating cells were weakly positive for OCT4 and positive for T, a marker for primitive streak/incipient mesoderm (Fig. 2c, d, Extended Data Fig. 2e, f). The continued expression of NANOG in EPI until late after implantation was a marked difference from the immediate repression of NANOG

¹Department of Anatomy and Cell Biology, Graduate School of Medicine, Kyoto University, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan. ²JST, ERATO, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan. ³Research Center for Animal Life Science, Shiga University of Medical Science, Seta-Tsukinowa-cho, Otsu, Shiga 520-2192, Japan. ⁴Center for iPS Cell Research and Application, Kyoto University, 53 Kawahara-cho, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan. ⁵Institute for Integrated Cell-Material Sciences, Kyoto University, Yoshida-Ushinomiya-cho, Sakyo-ku, Kyoto 606-8501, Japan. ⁶AMED-CREST, AMED, 1-7-1 Otemachi, Chiyoda-ku, Tokyo, 100-0004, Japan.

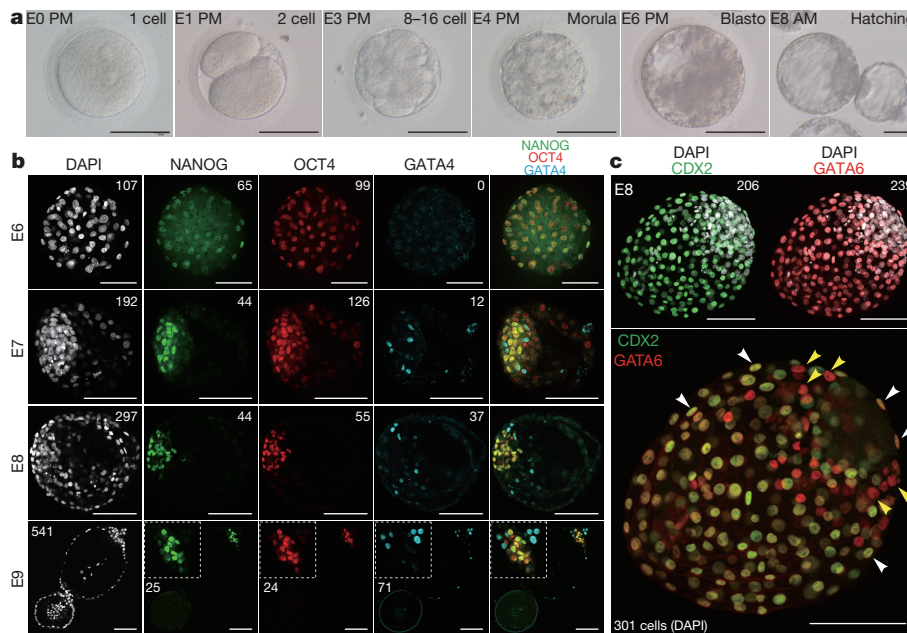


Figure 1 | Monkey pre-implantation development. **a**, Images of monkey pre-implantation embryos. Oocytes were fertilized around noon and the embryos were observed every day until E8 (embryos, $n = 68$). AM, 9:00; PM, 21:00. Embryos with more than 16 cells without blastocoel cavities, those with blastocoel cavities, and those with cells outside the zona pellucida were classified as morula, blastocysts and hatched, respectively.

after implantation in mice¹⁸. Extra-embryonic mesenchyme cells were positive for GATA4 and GATA6 (Extended Data Fig. 2f), consistent with their derivation from hypoblast/VE/YE^{15,17}.

Lineage delineation by transcriptome

We prepared single-cell cDNAs from pre- and post-implantation embryos using the single-cell mRNA 3-prime end sequencing (SC3-seq) method¹⁹ and screened all of the cDNAs for key markers by quantitative PCR (qPCR) (Extended Data Fig. 3). For pre-implantation cells (Extended Data Fig. 3b), at E6, *POU5F1* (gene encoding OCT4) was expressed ubiquitously, and *NANOG* exhibited more heterogeneous expression. With the six markers (*POU5F1*, *NANOG*, *CDX2*, *TFAP2C*, *GATA4* and *GATA6*), we could not discriminate trophoblast until E7. We detected low levels of *GATA4* in some cells as early as E6. At E7, *GATA4*^{hi} cells were negative for *NANOG* expression, suggesting hypoblast specification. At E8 and E9, the cells were classified into three types: *NANOG*^{hi}; *GATA4*[−] (EPI), *NANOG*^{lo/−}; *GATA4*^{hi} (hypoblast), and *NANOG*^{lo/−}; *GATA4*[−]; *GATA6/CDX2/TFAP2C*^{hi} (trophoblast), indicating a clear segregation of the three lineages.

For post-implantation embryos (Extended Data Fig. 3c), we generated cDNAs from cells expressing key pluripotency markers (*POU5F1*, *NANOG*, *SOX2*, *PRDM14*), representing EPI and its related lineages, as well as cDNAs from cells negative for such markers, but positive for *GATA4*, representing VE/YE and its related lineages. Notably, not only *NANOG*, but also *PRDM14* (refs 20, 21), continued to show high expression until E17. Some of the cells that were positive for pluripotency markers were positive for *T* as early as E13, and cells with expression of *T* continued to express *POU5F1*, but tended to repress *NANOG*, *SOX2*, and *PRDM14*, particularly after E16, indicating their mesodermal or endodermal identity. We isolated cells that were negative for the pluripotency markers and *GATA4*, but positive for *GATA2*, presumably representing post-implantation parietal trophoblast. We also isolated candidates for primordial germ cells (*SOX17*⁺*PRDM14*⁺*TFAP2C*⁺ and *SOX2*[−]), which will be analysed separately (data not shown).

We analysed the transcriptome of 390 cells (pre, 193; post, 197) (Extended Data Fig. 3, Supplementary Table 1) by SC3-seq, the

performance of which was better than or comparable to those of other single-cell RNA-seq methods^{19,22–24} (Extended Data Fig. 4). Unsupervised hierarchical clustering (UHC) classified all the cells into two large clusters, one consisting mainly of pre-implantation and the other only of post-implantation cells (Fig. 3). The principal component analysis (PCA) and a correlation analysis provided consistent outcomes (Extended Data Fig. 5a, b). On the basis of the markers and the cells' developmental stages, we annotated pre-implantation cells at E6 as either undifferentiated ICM or pre-implantation early trophoblast (preE-EPI), cells at E7 as either pre-implantation epiblasts (preE-EPI), preE-TE, pre-implantation late TE (preL-TE) or hypoblast, and cells at E8 and E9 as either pre-EPI, preL-TE, or hypoblast (Fig. 3, Extended Data Fig. 5a, b). PCA revealed a close relationship between ICM and preE-TE, a progressive transition from ICM to pre-EPI and from preE- to preL-TE, and a distinct property of hypoblast (Extended Data Fig. 5c). The differentially expressed genes among the annotated lineages are listed in Supplementary Table 2. Similarly, we annotated the post-implantation cells as the post-implantation early or late EPI (postE-EPI or postL-EPI), gastrulating cells 1, 2a and 2b (Gast1, 2a and 2b), VE/YE, and extra-embryonic mesenchyme (EXMC), and a distinct group of cells clustered with preL-TE as post-implantation parietal trophoblast (Fig. 3, Extended Data Fig. 5a, b).

POU5F1 and *NANOG* were highly expressed in post-EPI (postE-EPI and postL-EPI) and Gast1, whereas *SOX2* and *PRDM14* were down-regulated in a number of Gast1, and postL-EPI and Gast1, respectively (Extended Data Fig. 5d). *T* showed high expression in Gast1 and Gast2a as well as in some of post-EPI, but was low/negative in Gast2b, VE/YE, and extra-embryonic mesenchyme (Extended Data Fig. 5d). *GATA4* and *GATA6* were expressed in Gast1, Gast2a and Gast2b as well as in VE/YE and extra-embryonic mesenchyme, except that *GATA6* was negative in a majority of Gast1 (Extended Data Fig. 5d). We identified *FOXA1* as a gene specifically expressed in VE/YE but not in extra-embryonic mesenchyme, and reciprocally, *COL6A1* showed strong expression in extra-embryonic mesenchyme, but weak in VE/YE (Extended Data Fig. 5d, e). *T*, *MIXL1*, and *CDX2* were low/negative, but markers for epithelial-mesenchymal transition were high in extra-embryonic mesenchyme (Extended Data Fig. 5d). The genes highly expressed

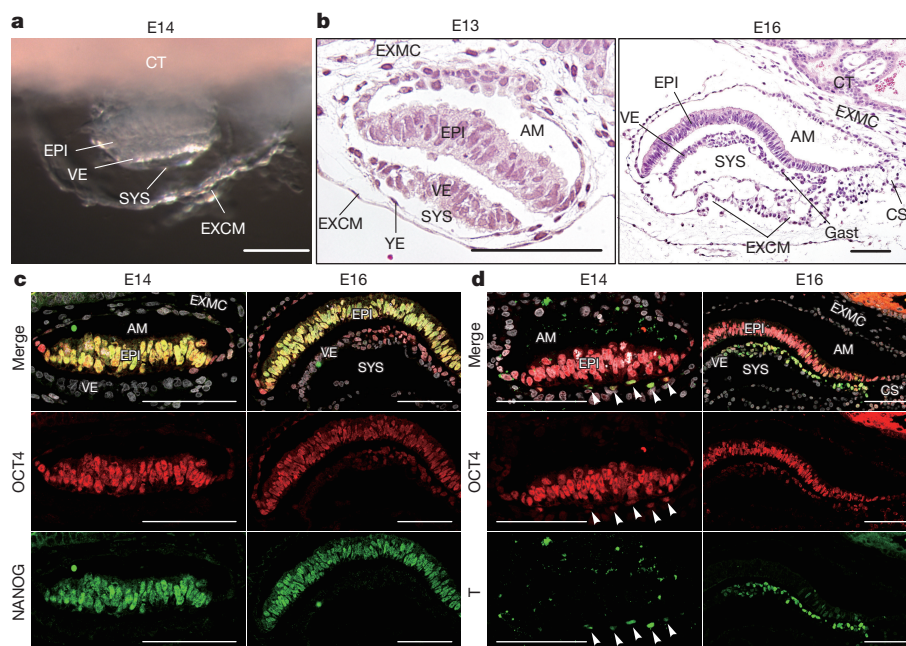


Figure 2 | Monkey early post-implantation development. **a**, A post-implantation embryo at E14 observed under a dissection microscopy. **b**, Haematoxylin and eosin staining of the sections of post-implantation embryos at E13 and E16. **c**, **d**, Expression of OCT4/NANOG (**c**), and OCT4/T (arrowheads) (**d**) in embryos at E14 and E16 (embryos $n = 2$, 2,

respectively). When recognizable, anterior is to the left. AM, amnion; CT, cytotrophoblast; CS*, future connecting stalk; EXMC, extra-embryonic mesenchyme; EXCM, exocoelomic membrane; Gast, gastrulating cells; SYS, secondary yolk sac; VE, visceral endoderm; YE, yolk-sac endoderm. Scale bars, 100 μ m.

in extra-embryonic mesenchyme were enriched with those bearing Gene Ontology terms such as 'extracellular matrix' (Extended Data Fig. 5f–h, Supplementary Table 2), and some of them exhibited specific/high expression in hypoblast (Extended Data Fig. 5g). Conversely, some of the genes that were highly expressed in hypoblast remained high in extra-embryonic mesenchyme (Extended Data Fig. 5i, j, Supplementary Table 2). Extra-embryonic mesenchyme exhibited transcriptional heterogeneity similar to postE-EPI (Extended Data Fig. 5k). Thus, trophoblast and hypoblast are specified by E6 and E7, respectively. EPI originates from ICM at E6 and through implantation, progressively acquires distinct properties, while generating gastrulating cells from around E13. Extra-embryonic mesenchyme cells appear to derive from hypoblast through epithelial–mesenchymal transition and generate abundant extracellular matrix (Extended Data Fig. 5l).

Transition of pluripotency properties

We focused on the transition of the properties of EPI. The expression of key genes associated with naive and primed pluripotency during

EPI development is shown in Extended Data Fig. 6a. The major changes occur during the ICM to pre-EPI and the pre-EPI to postE-EPI transitions with relatively large numbers of differentially expressed genes, whereas post-EPI stably maintains its properties (Extended Data Fig. 6b–d, Supplementary Table 2). Consequently, we found that post-EPI acquire genes associated, most prominently, with 'neuron differentiation/development', whereas they downregulate genes associated with 'regulation of cell proliferation' and 'mitochondrion/oxidative reduction' (Extended Data Fig. 6c, Supplementary Table 2). We confirmed robust expression of SOX11, a marker for neural tubes in mice²⁵, in E14 and E16 post-EPI (Extended Data Fig. 6e). The genes upregulated in Gast1 or Gast2a exhibited a marked enrichment for 'pattern specification process/embryonic morphogenesis' (Extended Data Fig. 6c, Supplementary Table 2).

To identify genes characteristic of the transition of EPI, we performed PCA of the EPI lineage, and determined the 776 genes with significantly positive or negative scores of the principal component 1/2 (PC1/2) loading (cyEPI ontogenic genes) (Fig. 4a, b, Supplementary Table 2).

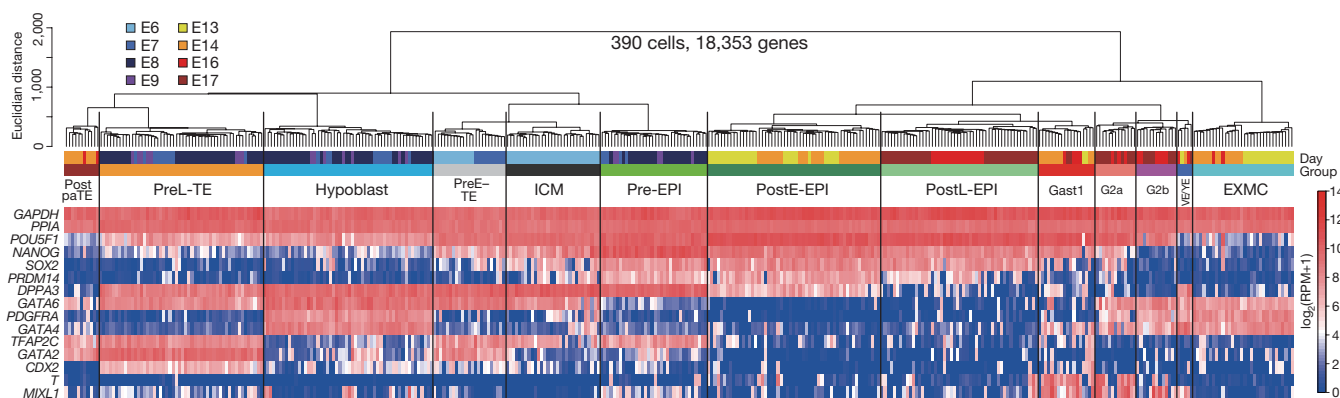


Figure 3 | Classification of key cell types by SC3-seq. Unsupervised hierarchical clustering (UHC) with all expressed genes (390 cells, 18,353 genes) and a heat map of the levels of selected marker genes. Colour bars under the dendrogram indicate embryonic days (top)

and cell types (bottom), respectively. G2a/G2b, Gast2a/Gast2b; pa, parietal; pre, pre-implantation; post, post-implantation; TE, trophoblast. Other abbreviations are as indicated in Fig. 2. The colour coding is as indicated.

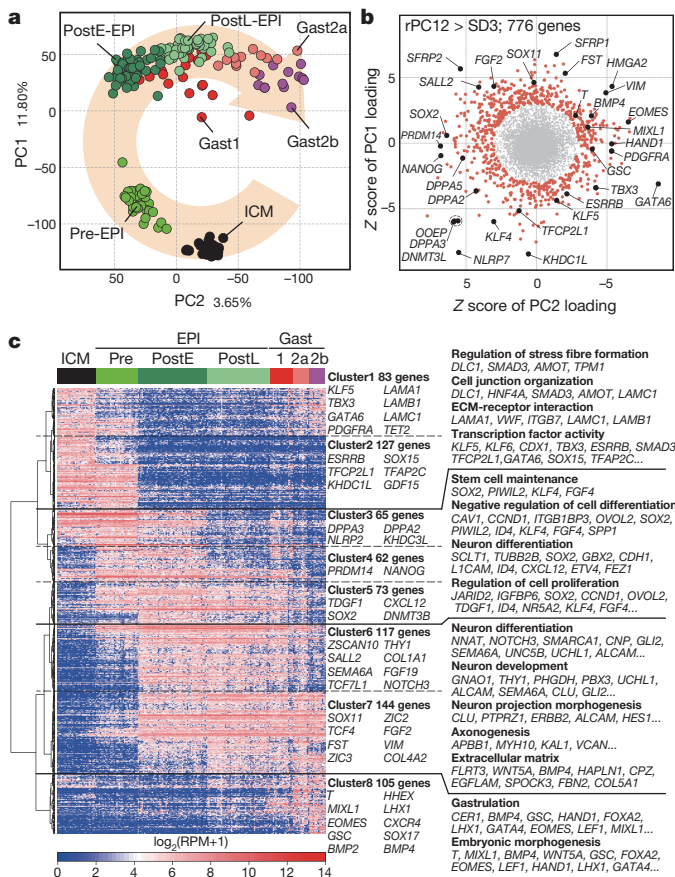


Figure 4 | Progressive transitions of the properties of cyEPI. a, PCA of the EPI lineage by all expressed genes among these groups (213 cells, 17,193 genes). b, Scatter plot of the normalized loading scores of PCA in a. Orange dots (776 genes: cyEPI ontogenic genes (Supplementary Table 2)) indicate genes that contributed highly to the PC1 and PC2 axes: more than 3 s.d. radius of PC1 and PC2 ($rPC12 > SD3$). Key genes are annotated. c, Heat map of the expression of cyEPI ontogenic genes. The genes were ordered by UHC, and eight clusters were defined according to the UHC dendrogram (left). Representative genes and key Gene Ontology enrichments are shown (Supplementary Table 2). RPM, reads per million mapped reads.

The UHC classified them into clusters exhibiting characteristic expression in the seven cell types (Fig. 4c). Clusters 1 and 2 were expressed in pre-implantation cells with cluster 1 being upregulated in *Gast2a/2b*, and were enriched with genes for cell morphology or extracellular matrix, and for ‘transcription factor activity’, including naive pluripotency genes (*KLF5*, *TBX3*, *ESRRB*, *TFCP2L1*, and *SOX15*) (Fig. 4c). Clusters 3 to 5 exhibited expression linking pre- and post-EPI, and their genes were linked with ‘stem cell maintenance’, including *KLF4*, *PRDM14*, *NANOG*, and *SOX2* (Fig. 4c). Clusters 6 and 7 showed expression mainly in post-implantation cells with cluster 6 being down-regulated in *Gast2b*, and exhibited a striking enrichment for ‘neuron differentiation’ (Fig. 4c). Genes expressed in post-E-EPI were typically expressed in *Gast1* (Fig. 4c). Cluster 8 was typically upregulated in *Gast* cells and enriched with genes for ‘gastrulation’ (Fig. 4c). Thus, cyEPI ontogenic genes distinguish the spectrum of pluripotency during EPI development.

Cynomolgus versus mouse EPI development

To compare cynomolgus and mouse EPI development, we analysed single-cell cDNA from EPI (*Pou5f1*⁺) and visceral endoderm (*Pou5f1*⁻*Gata4*⁺ at E5.5; *Pou5f1*⁻*Afp*⁺ at E6.5) of E5.5 and E6.5 mouse embryos by SC3-seq (Extended Data Fig. 7a, b, Supplementary Table 1). We also analysed the data for EPI, trophoblast, and primitive

endoderm of E4.5 embryos¹⁹. UHC classified these cells in a stage-dependent manner and PCA revealed a directional progression of EPI development, with E6.5 EPI exhibiting a scattered distribution, indicating their heterogeneity (Extended Data Fig. 7c, d). Around half of E6.5 EPI were highly positive for *T* (E6.5EPI-*T*^{hi}), and they were enriched with genes for ‘embryonic morphogenesis/pattern specification process’ (Extended Data Fig. 7e, f, Supplementary Table 2). mEPI changed their properties more rapidly and profoundly than cyEPI (Extended Data Fig. 7g). The genes upregulated during the E4.5–E5.5 transition were enriched for ‘apoptosis/programmed cell death’ and ‘regulation of transcription’, but unlike in cynomolgus monkeys, not for ‘neuron differentiation’, whereas those downregulated were enriched for ‘stem cell maintenance’ (*Il6st*, *Lifr*, *Nanog*, *Sox2*, *Esrrb*, *Klf4/5*, *Tbx3*, *Nr5a2*) (Extended Data Fig. 7g).

PCA revealed that the cynomolgus and mouse cells exhibited a separation along the PC1, which represents a major species difference (Extended Data Fig. 8a). We identified genes with significantly positive or negative PC1 scores, representing the cynomolgus and mouse genes, respectively, which were enriched for 'mitochondrion/oxidative reduction' and 'neuron projection morphogenesis' (cynomolgus genes), and 'ubiquitin mediated proteolysis/cell cycle/DNA repair' (mouse genes) (Extended Data Fig. 8b, Supplementary Table 2). The enrichment of cell-cycle related genes among mouse genes was consistent with the rapid proliferation of mouse embryonic cells. Notably, along the PC2 and 3 axes, the cynomolgus and mouse cells were plotted in a manner reflecting their developmental transitions (Extended Data Fig. 8a). We determined the genes with significantly positive or negative scores for PC2 and 3 loading, and we subtracted from them the genes that had significantly positive or negative scores for PC1 loading. A correlation analysis using the resulting gene set (monkey and mouse common EPI genes, 473 genes, Extended Data Fig. 8c, Supplementary Table 2) revealed that postE-EPI, postL-EPI, and *Gast1* exhibited the closest correlation with E5.5 EPI, whereas *Gast2a/2b* showed the closest correlation with E6.5EPI- T^{hi} (Extended Data Fig. 8d), defining an approximate developmental coordinate between the two species. Thus, cyEPI retain a property of pre-gastrulating mEPI long after implantation (~ 1 week).

To gain insight into a mechanism for the acquisition of the ‘neuron differentiation’ property in monkey post-EPI, we performed the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis for genes upregulated during the pre-EPI to postE-EPI transition, which revealed an enrichment of the NOTCH signalling pathway (Extended Data Fig. 8e–g). Constitutive activation of the NOTCH signalling in m/hPSCs or mouse embryos predisposes them for neuronal differentiation, with downregulation of the NODAL signalling^{26,27}. Consistently, *NODAL* was downregulated in post-EPI (Extended Data Fig. 8f). Contrastingly, the NOTCH signalling pathway was not upregulated in E5.5 mEPI, and *Nodal*, which prevents precocious neural differentiation²⁸, was strongly expressed in E5.5 mEPI (Extended Data Fig. 8e, f). Thus, the differential NOTCH signalling may create a difference between cynomolgus and mouse post-implantation EPI.

Developmental coordinate of pluripotency

We next sought to compare PSCs *in vitro* with the EPI lineage, and performed SC3-seq for cyESCs²⁹ (Extended Data Fig. 9a–c, Supplementary Table 1). cyESCs were clustered together with postE-EPI, postL-EPI, and Gast1, but were highly distinct from ICM and pre-EPI; note that one cyESC was clustered within postL-EPI (Extended Data Fig. 9d). Consistently, PCA plotted cyESCs closest to postL-EPI (Fig. 5a). When compared with pre-EPI, cyESCs up- and downregulated as many as 520 and 394 genes, respectively, whereas when compared with postL-EPI, cyESCs up- and downregulated only 170 and 26 genes, respectively (Fig. 5b, Supplementary Table 2). The genes upregulated in cyESCs against pre-EPI were enriched with those for ‘neuron differentiation’, and less significantly those for cell cycles

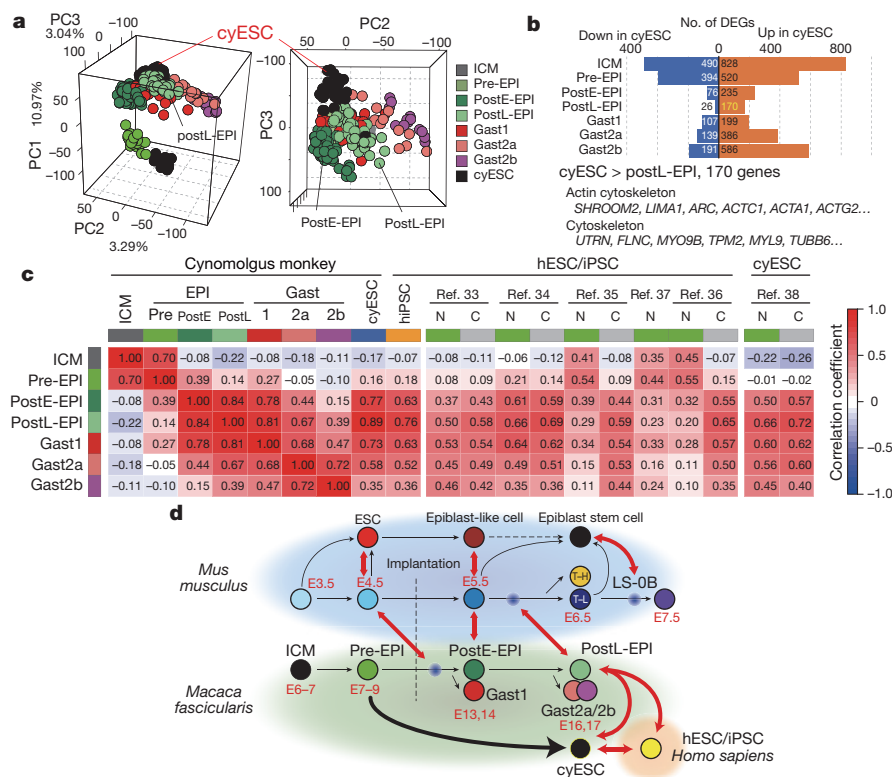


Figure 5 | Correlations among cells during cyEPI development and hPSCs/cyPSCs. **a**, PCA of monkey embryonic cells and cyESCs by all expressed genes among these groups (244 cells, 17,450 genes). The colour coding is as indicated. **b**, Differentially expressed genes between cells during the cyEPI development and cyESCs. Top, Orange and blue bars indicate the numbers of up- and downregulated genes, respectively, in the pair-wise comparisons indicated. Bottom, Representative Gene Ontology terms for genes upregulated in cyESC against postL-EPI (Supplementary

Table 2). **c**, Heat map of the correlation coefficients among the indicated cells including those reported by others^{33–38} based on the cyEPI ontogenic gene levels. The genes in common for all platforms were used (628 out of 776 genes). N, 'naive'; C, conventional. **d**, A model for a developmental coordinate of the spectrum of human, monkey and mouse pluripotency. Black arrows, direction of differentiation/derivation; red arrows, homologous relationships.

(Extended Data Fig. 9e, f). The genes upregulated in cyESCs against postL-EPI were enriched with those for 'actin cytoskeleton', suggesting an adaptation in culture (Fig. 5b, Supplementary Table 2). Thus, cyESCs bear properties corresponding to postL-EPI, which co-exists with gastrulating cells and is a precursor for gastrulation.

Contrastingly, mESCs with a 'ground state' of pluripotency^{30,31} were similar to E4.5 mEPI, whereas epiblast-like cells induced from mESCs³² were highly similar to E5.5 mEPI (Extended Data Fig. 10a–e); the number of differentially expressed genes between epiblast-like cells and E5.5 mEPI was much smaller than that between mESCs and E4.5 mEPI (Extended Data Fig. 10f, Supplementary Table 2). Cross-species comparison using the monkey and mouse common EPI genes revealed that cyESCs were correlated with E5.5 mEPI and epiblast-like cells, and reciprocally, epiblast-like cells were correlated with postL-EPI and cyESCs (Extended Data Fig. 10g, h). Interestingly, mESCs were closer to postE-EPI than to pre-EPI (Extended Data Fig. 10g, h).

We next examined the relationship between hiPSCs¹⁹ (Extended Data Fig. 9c) and the cyEPI lineage. As there were still considerable species differences between humans and cynomolgus monkeys, we used the cyEPI ontogenic genes. Remarkably, hiPSCs exhibited a profile highly similar to that of cyESCs, and accordingly, to that of postL-EPI (Fig. 5c, Extended Data Fig. 9g, h). We confirmed that human pre-implantation EPI (and also marmoset ICM (EPI and hypoblast))^{22–24} were similar to cynomolgus pre-EPI, but were distinct from post-EPI (Extended Data Fig. 9g, h), and interestingly, human pre-implantation EPI cultured for the establishment of hESCs acquired a similarity to monkey post-EPI at passage 0 (Extended Data Fig. 9g, h). We examined the properties of hPSCs and cyPSCs that have been previously reported to show naive pluripotency^{33–38}. Notably, hESCs reported in refs 35–37

exhibited the closest correlation with pre-EPI, whereas hESCs in ref. 34 and cyESCs in ref. 38 remained essentially unchanged from their original states and were similar to postL-EPI, and hESCs reported in ref. 33 showed the closest correlation with Gast1, reflecting their expression of genes for gastrulation (Fig. 5c, Extended Data Fig. 9g).

Discussion

Through the systematic use of SC3-seq, we have established a transcriptional foundation for the origin and development of cyEPI, defining a developmental coordinate of pluripotency among mice, monkeys, and humans (Fig. 5d). After implantation, while generating gastrulating cells, cyEPI maintains its transcriptional property stably for a prolonged period (~1 week), suggesting that cyEPI could be a source for the cells of the same lineages for an extended period. The finding that cy/hPSCs bear transcriptional properties similar to postL-EPI and, to a slightly lesser extent, to postE-EPI, is consistent with the fact that hPSCs are induced into cells in the three germ layers and primordial germ cell-like cells^{39,40}, although cy/hPSCs lack naive pluripotency and show line-dependent differentiation biases^{9,41}. cyEPI ontogenic genes (Fig. 4b, c) were highly instructive for defining the properties of hPSCs (Fig. 5c); considering the controversial state of naive hPSCs¹⁰ and a higher similarity of mESCs to postE-EPI than to pre-EPI (Extended Data Fig. 10g, h), exploring a condition that provides hPSCs with a property more similar to postE-EPI might be one strategy for improving the potentials of hPSCs. It should also be interesting to compare monkey post-EPI with the human epiblast after 'implantation culture' of human blastocysts^{42,43}. The future challenges will include exploring the mechanism for monkey lineage specification as well as for the maturation of cyEPI, and performing more comprehensive analysis for monkey gastrulation. Such investigation will lead to a better

strategy for controlling the properties of hPSCs and for generating cells of interest from hPSCs.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 November 2015; accepted 12 July 2016.

Published online 24 August 2016.

- Gardner, R. L. & Rossant, J. Investigation of the fate of 4–5 day post-coitum mouse inner cell mass cells by blastocyst injection. *J. Embryol. Exp. Morphol.* **52**, 141–152 (1979).
- Lawson, K. A., Meneses, J. J. & Pedersen, R. A. Clonal analysis of epiblast fate during germ layer formation in the mouse embryo. *Development* **113**, 891–911 (1991).
- Evans, M. J. & Kaufman, M. H. Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292**, 154–156 (1981).
- Thomson, J. A. *et al.* Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145–1147 (1998).
- Tesar, P. J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448**, 196–199 (2007).
- Brons, I. G. *et al.* Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature* **448**, 191–195 (2007).
- Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
- Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).
- Takahashi, K. & Yamanaka, S. Induced pluripotent stem cells in medicine and biology. *Development* **140**, 2457–2461 (2013).
- Rossant, J. Mouse and human blastocyst-derived stem cells: vive les differences. *Development* **142**, 9–12 (2015).
- Nichols, J. & Smith, A. Naive and primed pluripotent states. *Cell Stem Cell* **4**, 487–492 (2009).
- Yamasaki, J. *et al.* Vitro fertilization and transfer of cynomolgus monkey (*Macaca fascicularis*) embryos fertilized by intracytoplasmic sperm injection. *Theriogenology* **76**, 33–38 (2011).
- Niakan, K. K. & Eggan, K. Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. *Dev. Biol.* **375**, 54–64 (2013).
- Heuser, C. H. & Streeter, G. L. Development of the macaque embryo. *Contrib. Embryol.* **29**, 15–55 (1941).
- Lockett, W. P. Origin and differentiation of the yolk sac and extraembryonic mesoderm in presomite human and rhesus monkey embryos. *Am. J. Anat.* **152**, 59–97 (1978).
- Enders, A. C., Schlafke, S. & Hendrickx, A. G. Differentiation of the embryonic disc, amnion, and yolk sac in the rhesus monkey. *Am. J. Anat.* **177**, 161–185 (1986).
- Enders, A. C. & King, B. F. Formation and differentiation of extraembryonic mesoderm in the rhesus monkey. *Am. J. Anat.* **181**, 327–340 (1988).
- Acampora, D., Di Giovannantonio, L. G. & Simeone, A. Otx2 is an intrinsic determinant of the embryonic stem cell state and is required for transition to a stable epiblast stem cell condition. *Development* **140**, 43–55 (2013).
- Nakamura, T. *et al.* SC3-seq: a method for highly parallel and quantitative measurement of single-cell gene expression. *Nucleic Acids Res.* **43**, e60 (2015).
- Yamaji, M. *et al.* Critical function of *Prdm14* for the establishment of the germ cell lineage in mice. *Nat. Genet.* **40**, 1016–1022 (2008).
- Yamaji, M. *et al.* PRDM14 ensures naive pluripotency through dual regulation of signaling and epigenetic pathways in mouse embryonic stem cells. *Cell Stem Cell* **12**, 368–382 (2013).
- Yan, L. *et al.* Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1131–1139 (2013).
- Blakeley, P. *et al.* Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development* **142**, 3151–3165 (2015).
- Boroviak, T. *et al.* Lineage-Specific Profiling Delineates the Emergence and Progression of Naive Pluripotency in Mammalian Embryogenesis. *Dev. Cell* **35**, 366–382 (2015).
- Lin, L. *et al.* Sox11 regulates survival and axonal growth of embryonic sensory neurons. *Dev. Dyn.* **240**, 52–64 (2011).
- Lowell, S., Benchoua, A., Heavey, B. & Smith, A. G. Notch promotes neural lineage entry by pluripotent embryonic stem cells. *PLoS Biol.* **4**, e121 (2006).
- Souilh, C. *et al.* NOTCH activation interferes with cell fate specification in the gastrulating mouse embryo. *Development* **142**, 3649–3660 (2015).
- Camus, A., Perea-Gomez, A., Moreau, A. & Collignon, J. Absence of Nodal signaling promotes precocious neural differentiation in the mouse embryo. *Dev. Biol.* **295**, 743–755 (2006).
- Suemori, H. *et al.* Establishment of embryonic stem cell lines from cynomolgus monkey blastocysts produced by IVF or ICSI. *Dev. Dyn.* **222**, 273–279 (2001).
- Ying, Q. L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519–523 (2008).
- Boroviak, T., Loos, R., Bertone, P., Smith, A. & Nichols, J. The ability of inner-cell-mass cells to self-renew as embryonic stem cells is acquired following epiblast specification. *Nat. Cell Biol.* **16**, 516–528 (2014).
- Hayashi, K., Ohta, H., Kurimoto, K., Aramaki, S. & Saitou, M. Reconstitution of the mouse germ cell specification pathway in culture by pluripotent stem cells. *Cell* **146**, 519–532 (2011).
- Gafni, O. *et al.* Derivation of novel human ground state naive pluripotent stem cells. *Nature* **504**, 282–286 (2013).
- Chan, Y. S. *et al.* Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast. *Cell Stem Cell* **13**, 663–675 (2013).
- Theunissen, T. W. *et al.* Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. *Cell Stem Cell* **15**, 471–487 (2014).
- Takahashi, Y. *et al.* Resetting transcription factor control circuitry toward ground-state pluripotency in human. *Cell* **158**, 1254–1269 (2014).
- Guo, G. *et al.* Naive pluripotent stem cells derived directly from isolated cells of the human inner cell mass. *Stem Cell Reports* **6**, 437–446 (2016).
- Chen, Y. *et al.* Generation of cynomolgus monkey chimeric fetuses using embryonic stem cells. *Cell Stem Cell* **17**, 116–124 (2015).
- Sasaki, K. *et al.* Robust *in vitro* induction of human germ cell fate from pluripotent stem cells. *Cell Stem Cell* **17**, 178–194 (2015).
- Irie, N. *et al.* SOX17 is a critical specifier of human primordial germ cell fate. *Cell* **160**, 253–268 (2015).
- Kajiwar, M. *et al.* Donor-dependent variations in hepatic differentiation from human-induced pluripotent stem cells. *Proc. Natl Acad. Sci. USA* **109**, 12538–12543 (2012).
- Degincerti, A. *et al.* Self-organization of the *in vitro* attached human embryo. *Nature* **533**, 251–254 (2016).
- Shahbazi, M. N. *et al.* Self-organization of the human embryo in the absence of maternal tissues. *Nat. Cell Biol.* **18**, 700–708 (2016).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported in part by a Grant-in-Aid from MEXT and by JST-ERATO. We thank Y. Nagai, R. Kabata, N. Konishi, Y. Sakaguchi, M. Kasawaki, T. Sato, M. Kabata, J. Matsushita, and M. Matsutani for their technical assistance. We are grateful to the Center for Anatomical, Pathological and Forensic Medical Researches, Kyoto University, for histology, to H. Suemori for CMK6/9, to M. Ema for encouragement, and to the animal care staff at Research Center for Animal Life Science, Shiga University of Medical Science for assistance.

Author Contributions T.N. and M.S. conceived the project, designed the experiments and wrote the manuscript. T.N. conducted all the experiments and analysed the data, I.O. helped with single-cell RNA seq experiments, K.S., C.I., H.T., Y.S., and S.N. assisted the isolation of cynomolgus monkey embryos, and Y.Y. and T.Y. assisted the analysis of single-cell RNA-seq data.

Author Information The SC3-seq data generated in this study have been deposited at Gene Expression Omnibus (GEO) database under accession number GSE74767. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.S. (saitou@anat2.med.kyoto-u.ac.jp).

Reviewer Information Nature thanks T. Li, K. Niakan and H. Niwa for their contribution to the peer review of this work.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Animals. Experimental procedures using cynomolgus monkeys were approved by the Animal Care and Use Committee of Shiga University of Medical Science. The procedures in cynomolgus monkeys for oocyte collection, intra-cytoplasmic sperm injection, pre-implantation embryo culture, and transfer of pre-implantation embryos into foster mothers were performed as described previously with minor modifications¹². For super-ovulation, ovarian stimulation with follicle-stimulating hormone (Gonapure, ASKA) was performed by embedding an implantable and programmable micro-fusion device (iPRECIO; Primetech Corporation) subcutaneously. The day when the intra-cytoplasmic sperm injection was performed was designated as embryonic day (E) 0. Embryos with more than 16 cells without blastocoel cavities, those with blastocoel cavities, and those with cells outside the zona pellucida were classified as morula, blastocysts, and hatched, respectively. The progression of the pre-implantation development of cynomolgus monkeys was highly similar to that of rhesus monkeys⁴⁴, but was somewhat slower than that of humans⁴⁵. For embryo transfer, 4 to 5 two-cell to blastocyst-stage embryos were selected and transferred into an appropriate recipient female. For the detection of pregnancy of early post-implantation embryos, implanted embryos were monitored by ultrasound scanning around E13 and the implanted uterus was surgically removed and bisected for the isolation of embryos.

Experimental procedures using mice were performed under the ethical guidelines of Kyoto University. For isolating mouse embryos, C57BL/6 mice were mated, and noon of the day when a copulation plug was identified was designated as E0.5. We did not determine the sex of embryos.

Immunofluorescence analysis. For monkey pre-implantation embryos, whole embryos were fixed in 4% paraformaldehyde in PBS (N26126-25; Nacalai Tesque) for 20 min at room temperature, washed in 2% BSA/PBS and incubated in the permeabilization solution (0.5% Triton X (T9284; Sigma-Aldrich)/1.0% BSA (A4503; Sigma-Aldrich)/PBS (20012-027; Thermo Fisher Scientific)) for 20 min at room temperature. After washing twice with 2% BSA/PBS, the embryos were incubated with primary antibodies in 2% BSA/PBS overnight at 4°C, washed three times with 2% BSA/PBS, incubated with secondary antibodies and 4,6-diamidino-2-phenylindole (DAPI) in 2% BSA/PBS for 1 h at room temperature, washed three times with 2% BSA/PBS, and mounted in VECTASHIELD mounting medium (H-1000; Vector Laboratories). Image data were captured and processed by a confocal microscope (Olympus FV1000 or Zeiss LSM780).

For monkey post-implantation embryos, implantation sites were dissected out from the uterus and fixed in 10% formalin (37152-51; Nacalai Tesque) overnight at 4°C. The samples were embedded in paraffin and sectioned at a thickness of 2–4 µm. Each slide was treated with HistVT ONE (06380-05; Nacalai Tesque) according to the manufacturer's instructions and incubated in blocking solution (2% BSA/PBS). For primary antibody reaction, sections were incubated in the blocking solution with primary antibodies at 4°C overnight. After washing six times with PBS, sections were incubated in the blocking solution with secondary antibodies and DAPI for 1 h at room temperature, washed four times with PBS-T (0.05% Tween 20 (P9416; Sigma-Aldrich) in PBS), washed twice with PBS, and mounted in VECTASHIELD mounting medium (H-1000; Vector Laboratories). Image data were captured and processed by a confocal microscope (Olympus FV1000 or Zeiss LSM780). After acquisition of the immunofluorescence image, sections were re-stained with haematoxylin and eosin for histological analysis. For Fig. 2c, d, Extended Data Fig. 2c, several images were acquired for one large area, and merged using the Photomerge function of Photoshop CC (Adobe Systems). All the antibodies used in this study are listed in Supplementary Table 1 along with the information on dilution ratios.

Cell culture. CMK6 and CMK9 were gifts from H. Suemori²⁹. For cultivation on feeders, they were cultured with conventional hESC medium (DMEM/F12 (D6421; Sigma-Aldrich) supplemented with 20% (vol/vol) of KSR (10828-028; Thermo Fisher Scientific), 1 mM of sodium pyruvate (11360-070; Thermo Fisher Scientific), 2 mM of GlutaMax (35050-061; Thermo Fisher Scientific), 0.1 mM of non-essential amino acids (11143-050; Thermo Fisher Scientific), 0.1 mM of 2-mercaptoethanol (M3148; Sigma-Aldrich), 1,000 U ml⁻¹ of ESGRO mouse LIF (ESG1107; Millipore), and 4 ng ml⁻¹ of recombinant human bFGF (060-04543; Wako Pure Chemical Industries)) on mouse embryonic feeders (MEFs). For feeder-free cultivation, cyESCs were cultured under the same condition as hiPSCs, as described previously^{39,46}. The cultivation of mESCs and the induction of day 2 epiblast-like cells were performed as described previously³². All of the cell lines used in this study have been tested for mycoplasma contamination by MycoAlert (LT07-118; Lonza Japan), according to the manufacturer's instructions.

Isolation of single cells for single-cell cDNA preparation. For monkey pre-implantation embryos, a whole embryo was incubated with 0.25% trypsin/PBS (T4799; Sigma-Aldrich) for around 10 min at 37°C, then dissociated into single cells by repeated pipetting, and dispersed in 0.1 mg ml⁻¹ of PVA/PBS (P8136; Sigma-Aldrich) for preparation of single-cell cDNAs.

For monkey post-implantation embryos, the implantation site was dissected out from the uterus and the embryonic fragment containing the epiblast (EPI), amnion, hypoblast, and yolk-sac endoderm was isolated manually. The fragment was incubated with 0.25% trypsin/PBS for around 10 min at 37°C, then dissociated into single cells by repeated pipetting, and dispersed in 0.1 mg ml⁻¹ of PVA/PBS.

For mouse E5.5 and E6.5 embryos, the embryos were dissected out from decidua and the extra-embryonic ectoderm was removed manually. The EPI/visceral endoderm were incubated in 0.25% Pancreatin (P3292; Sigma-Aldrich)/0.5% Trypsin/Polyvinylpyrrolidone (PVP40; Sigma-Aldrich), and EPI and visceral endoderm were separated by mild pipetting. For E5.5 embryos, EPI and visceral endoderm were incubated with 0.25% trypsin/PBS separately for around 10 min at 37°C. For E6.5 EPI, their proximal parts were dissected out manually, and the proximal EPI and whole visceral endoderm were incubated with 0.25% trypsin/PBS (Sigma-Aldrich) separately for around 10 min at 37°C. The proximal EPI and visceral endoderm were dissociated into single cells by repeated pipetting, and dispersed in 0.1 mg ml⁻¹ of PVA/PBS (Sigma-Aldrich).

For cyESCs, cells were first detached as clumps with CTK solution (0.25% of trypsin (15090-046; Thermo Fisher Scientific), 0.1 mg ml⁻¹ of collagenase IV (17104-019; Thermo Fisher Scientific), and 1 mM of CaCl₂ (06729-55; Nacalai Tesque)), incubated in TrypLE Select (12563029; Thermo Fisher Scientific) for around 10 min at 37°C, and dispersed into single cells in 1% (vol/vol) KSR/PBS containing 10 µM of the ROCK inhibitor Y-27632 (257-00511; Wako Pure Chemical Industries). Cells under feeder-free condition were directly incubated in TrypLE Select for around 5 min at 37°C, and dispersed into single cells in 1% (vol/vol) KSR/PBS containing 10 µM of the ROCK inhibitor Y-27632.

For mESCs and epiblast-like cells, cells were incubated in TrypLE Select for around 5 min at 37°C, and dispersed into single cells in 1% (vol/vol) KSR/PBS.

Single-cell cDNA preparation and transcriptome analysis by SC3-seq. cDNA synthesis and amplification from isolated single cells were performed essentially as described previously^{19,47,48}. Two types of spike-in RNAs—that is, the four *Bacillus subtilis* mRNAs, *lys*, *dap*, *phe* and *thr*, used in ref. 47 and 48, and the mRNAs developed by the External RNA Controls Consortium (ERCC; Life Technologies (4456740))—were used as described in ref. 19.

Before the construction of the SC3-seq library, the quality of the amplified cDNAs was evaluated by examining the C_t values of the qPCR of several endogenous genes, and by examining the distribution of the lengths of cDNA fragments using a LabChip GX (CLS760672; Perkin Elmer) or Bioanalyzer 2100 (5067-4626; Agilent Technologies) system. qPCR was performed using Power SYBR Green PCR Master Mix (4367659; Life Technologies) with a CFX384 real-time qPCR system (Bio-Rad, Hercules, CA) according to the manufacturer's instructions. The primer sequences are listed in Supplementary Table 1. Most of the primer sets were designed using Primer-Blast (NCBI) within a distance of 500 base pairs (bp) from the transcription termination sites (TTSs).

SC3-seq libraries of quality checked cDNAs were constructed as described previously¹⁹. The quality and quantity of the constructed libraries were evaluated by using a LabChip GX or Bioanalyzer 2100 system, a Qubit dsDNA HS assay kit (Q32851; Life Technologies), and a SOLiD Library TaqMan Quantitation kit (4449639; Life Technologies). The clonal amplification of the libraries on beads by emulsion PCR (emPCR) was performed using a SOLiD EZ Bead System (4449639; Life Technologies) at the E120 scale according to the manufacturer's instruction. The resulting bead libraries were loaded onto flowchips and sequenced for 50 bp and 5 bp barcode plus Exact Call Chemistry (ECC) on an SOLiD 5500XL system (4449639; Life Technologies).

Mapping reads of RNA-seq and conversion to gene expression levels. The genome sequence (GRCm38/mm10 for mice, GRCh37/hg19 for humans, and MacFas5.0 for cynomolgus monkeys (*Macaca fascicularis*)) and the transcript annotation (GRCm38 for mice, GRCh37 for humans, and MacFas5.0 for cynomolgus monkeys) were obtained from the NCBI ftp site at http://ftp-trace.ncbi.nlm.nih.gov/genomes/Macaca_fascicularis. SC3-seq reads only the 3' end of transcripts, so that the expression levels were calculated as genes (Entrez genes) but not mRNAs. Read trimming, mapping and estimation of expression levels were performed as described previously^{19,39}.

For mapping of the full-length RNA-seq data obtained externally, reads of data from^{22–24,33–38,49} were mapped onto the human (hg19), cynomolgus (MacFas5.0), marmoset (calJac3), or mouse (mm10) genome, using TopHat v2.0.11 (ref. 50), respectively. Mapped data were converted into expression levels using cufflinks

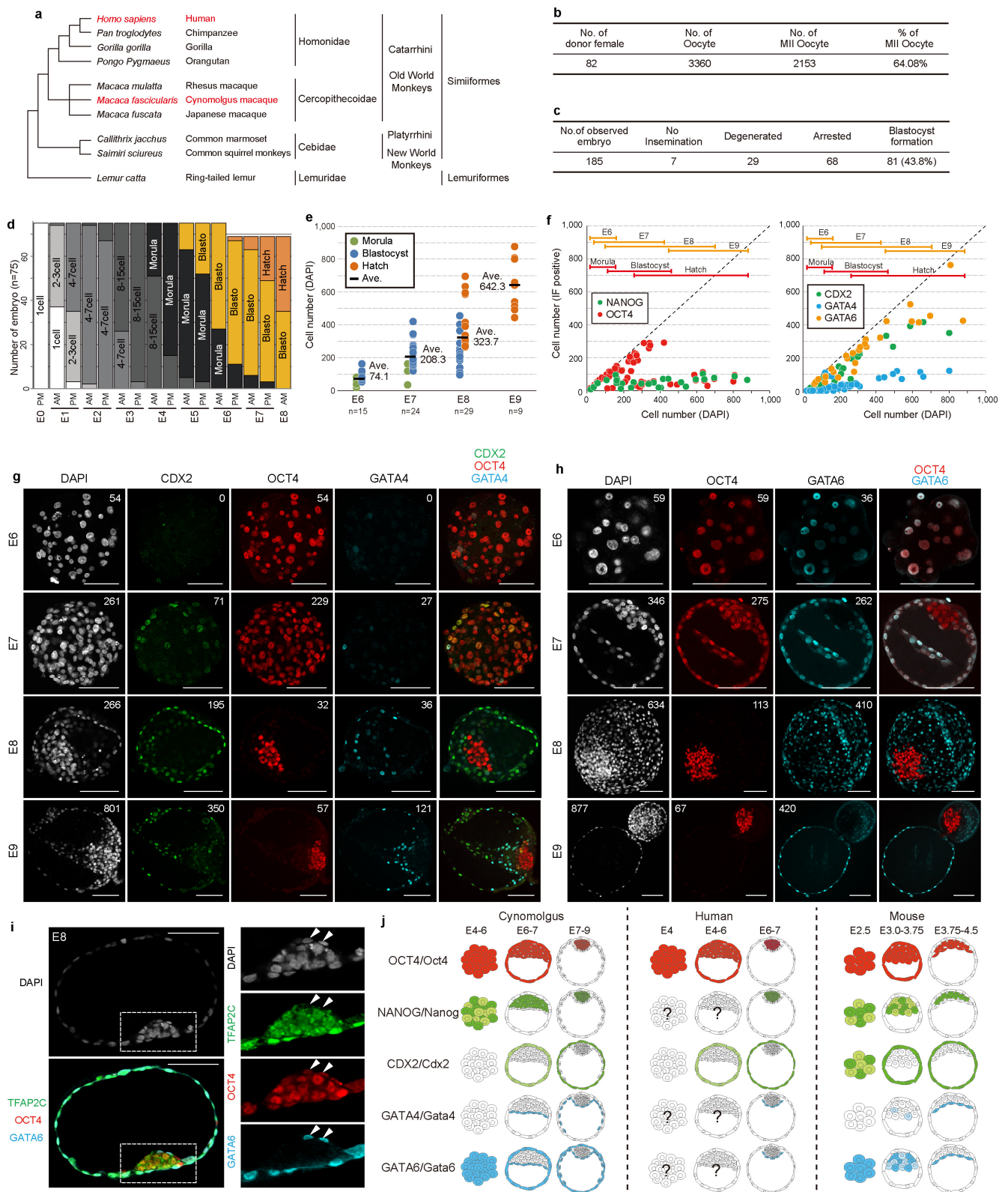
v.2.2.0 (ref. 51) with the ‘-compatible-hits-norm -library-type fr-unstranded-max-mle-iterations 50000’ options.

Data analysis of the SC3-seq. Data analysis was performed using R software version 3.1.1 with the gplots (ver. 2.16.0), qvalue (ver. 1.40.0), rgl (ver. 0.95.1201), vioplot (ver. 0.2), and genefilter (ver. 1.48.1) packages, and EXCEL (Microsoft), as described previously¹⁹. All the analyses of expression data were performed using $\log_2(\text{RPM}+1)$ values. We defined ‘all expressed genes’ as genes whose $\log_2(\text{RPM}+1)$ values were >4 (greater than ~ 10 –20 copies per cell, which is a lower limit for reliable and reproducible detection by our single-cell cDNA amplification method^{19,47}) in at least one sample. Unsupervised hierarchical clustering (UHC) was performed using the hclust function with Euclidian distances and Ward’s method (ward.D2). The principal component analysis (PCA) was performed using the prcomp function without scaling. To identify differentially expressed genes (DEGs) among multi-groups, the kruskal.test function for the Kruskal–Wallis test and the q value function were used for the calculation of the P value and false discovery ratio (FDR)⁵², respectively. The DEGs were defined as the genes exhibiting more than fourfold changes between the samples ($\text{FDR} < 0.01$), and the mean of the expression level of the group was $>\log_2(\text{RPM}+1) = 4$. For the Gene Ontology and Kyoto Encyclopedia of Genes and Genomes⁵³ (KEGG) analyses using the DAVID web tool⁵⁴, since the annotation of *Macaca fascicularis* genes was relatively incomplete, human annotation corresponding to that of cynomolgus monkeys was used. For this purpose, a one-to-one correspondence table of genes was made by genomic coordinate comparison using the LiftOver tool, as described previously³⁹ (Supplementary Table 2).

Comparison of gene expression between cynomolgus monkeys and humans, and between cynomolgus monkeys and mice. For comparison of the gene expression between cynomolgus monkeys and humans, common genes listed in the cynomolgus monkeys–humans one-to-one annotation table were used. Specifically, there are 28,551 genes in MacFas5.0 and 22,577 genes in GRCh37, with 17,542 genes in common between the two (Supplementary Table 2). For a comparison between cynomolgus monkeys and mice, first, a humans–mice annotation list was generated, and then a cynomolgus monkeys–humans list and mice–humans list were combined using human gene identifiers. There are 24,216 genes in GRCm38 and 15,933 genes in the mice–humans combined list, and consequently, 15,220 genes in the cynomolgus monkeys–humans–mice gene list (Supplementary Table 2). **Analysis of published expression data for human PSCs.** Expression levels of RNA-seq data were calculated as described above, and those of microarray data^{33,35} were obtained from a series matrix sheet in the GEO repository (NCBI). For data processing, expression levels of RNA-seq data were transformed into $\log_2(\text{fragments per kilobase of exon per million mapped sequence reads (FPKM)} + 0.1)$, and those of microarray data were transformed into $\log_2(\text{intensity})$. For comparison of microarray data to the SC3-seq data, the highest intensity probes were used for genes with multiple probes.

Accession numbers. Accession numbers for the data generated in this study and for the published data used in this study are as follows. The SC3-seq data in this study: GSE74767; those of mouse E4.5 cells and human iPSCs¹⁹: GSE63266; the transcriptome data for ref. 22, GSE36552; ref. 23, GSE66507; ref. 33, GSE46872; ref. 34, E-MTAB-2031; ref. 36, E-MTAB-2857; ref. 37, E-MTAB-4461; ref. 35: GSE59430; ref. 24: E-MTAB-2958, E-MTAB-2959; ref. 38: GSE69708; and ref. 49: GSE45916. The samples used are listed in Supplementary Table 1.

44. Wolf, D. P. *et al.* Use of assisted reproductive technologies in the propagation of rhesus macaque offspring. *Biol. Reprod.* **71**, 486–493 (2004).
45. Wong, C. C. *et al.* Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage. *Nat. Biotechnol.* **28**, 1115–1121 (2010).
46. Nakagawa, M. *et al.* A novel efficient feeder-free culture system for the derivation of human induced pluripotent stem cells. *Sci. Rep.* **4**, 3594 (2014).
47. Kurimoto, K. *et al.* An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Res.* **34**, e42 (2006).
48. Kurimoto, K., Yabuta, Y., Ohinata, Y. & Saitou, M. Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis. *Nat. Protocols* **2**, 739–752 (2007).
49. Ohta, S., Nishida, E., Yamanaka, S. & Yamamoto, T. Global splicing pattern reversion during somatic cell reprogramming. *Cell Reports* **5**, 357–366 (2013).
50. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
51. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protocols* **7**, 562–578 (2012).
52. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
53. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
54. Huang, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols* **4**, 44–57 (2009).
55. Perelman, P. *et al.* A molecular phylogeny of living primates. *PLoS Genet.* **7**, e1001342 (2011).
56. Dietrich, J. E. & Hiiragi, T. Stochastic patterning in the mouse pre-implantation embryo. *Development* **134**, 4219–4231 (2007).
57. Roode, M. *et al.* Human hypoblast formation is not dependent on FGF signalling. *Dev. Biol.* **361**, 358–363 (2012).
58. Xenopoulos, P., Kang, M., Puliafito, A., Di Talia, S. & Hadjantonakis, A. K. Heterogeneities in Nanog Expression Drive Stable Commitment to Pluripotency in the Mouse Blastocyst. *Cell Reports* **10**, 1508–1520 (2015).
59. Schrode, N., Saiz, N., Di Talia, S. & Hadjantonakis, A. K. GATA6 levels modulate primitive endoderm cell fate choice and timing in the mouse blastocyst. *Dev. Cell* **29**, 454–467 (2014).

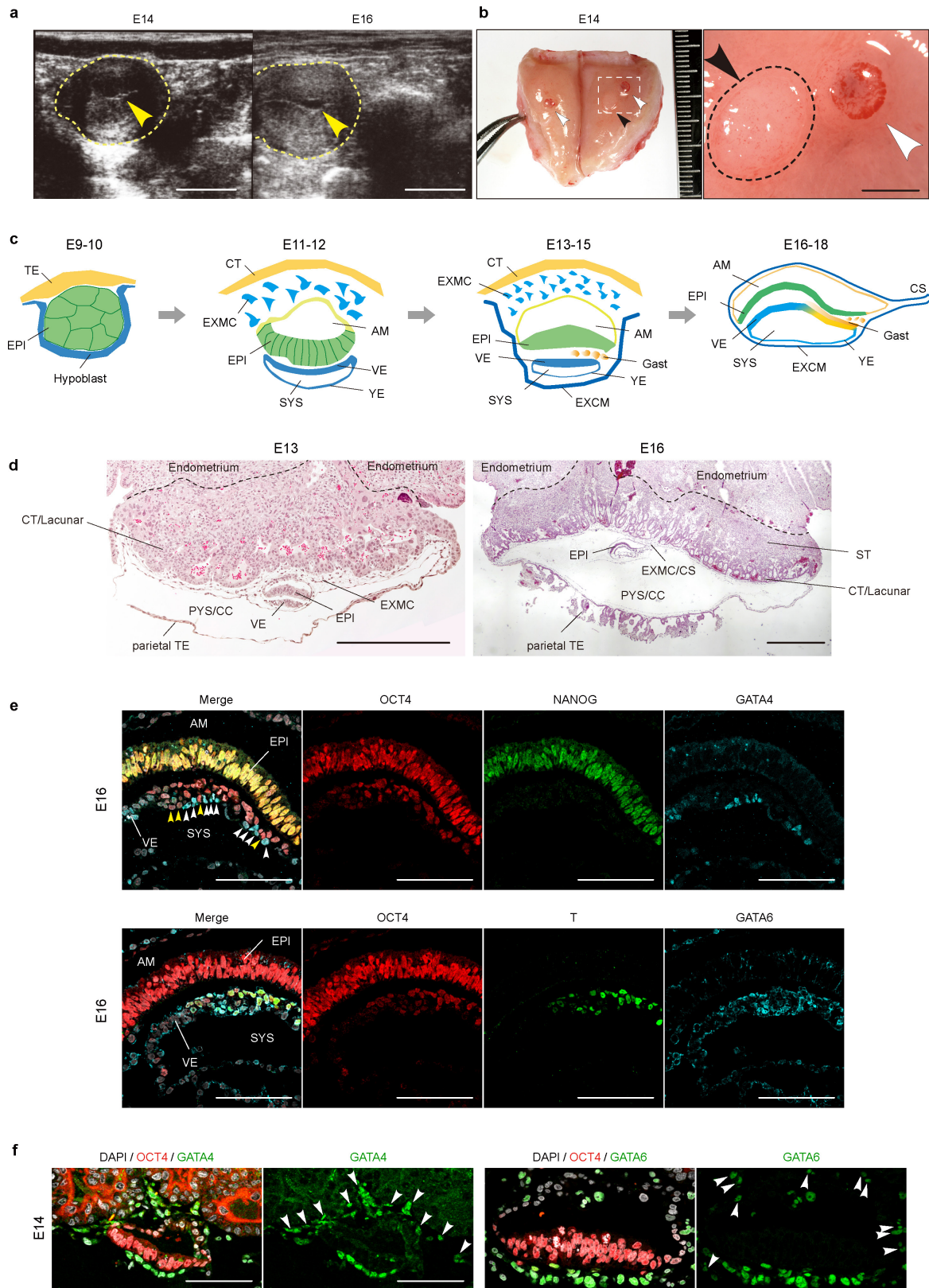


Extended Data Figure 1 | See next page for caption.

Extended Data Figure 1 | Summary of monkey pre-implantation development.

a, A phylogenetic tree of primates⁵⁵. *Cynomolgus*, rhesus, and Japanese monkeys are members of macaques of *Ceropithecoidae*, which are classified as old-world monkeys. **b**, Summary of super-ovulation and oocyte collection for this study. **c**, Summary of monkey pre-implantation development. No insemination, embryos without cleavage; degenerated, degenerated embryos; arrested, embryos that failed to form a blastocoel; blastocyst formation, embryos with blastocoel formation by E8. **d**, Developmental progression of monkey pre-implantation embryos. Embryos with more than 16 cells without blastocoel cavities, those with blastocoel cavities, and those with cells outside the zona pellucida were classified as morula, blastocysts, and hatched, respectively. **e**, The cell numbers (counted by DAPI) of pre-implantation embryos from E6 to E9. The cells with degenerated nuclei were excluded. The colour coding is as indicated. **f**, Scatter plots of the cell numbers that were positive for each marker (*y* axis, the colour coding indicated) against the whole-cell numbers (*x* axis). Each plot indicates the numbers in one embryo. The orange and red bars indicate the range of embryonic days and developmental stages, respectively. **g**, Expression

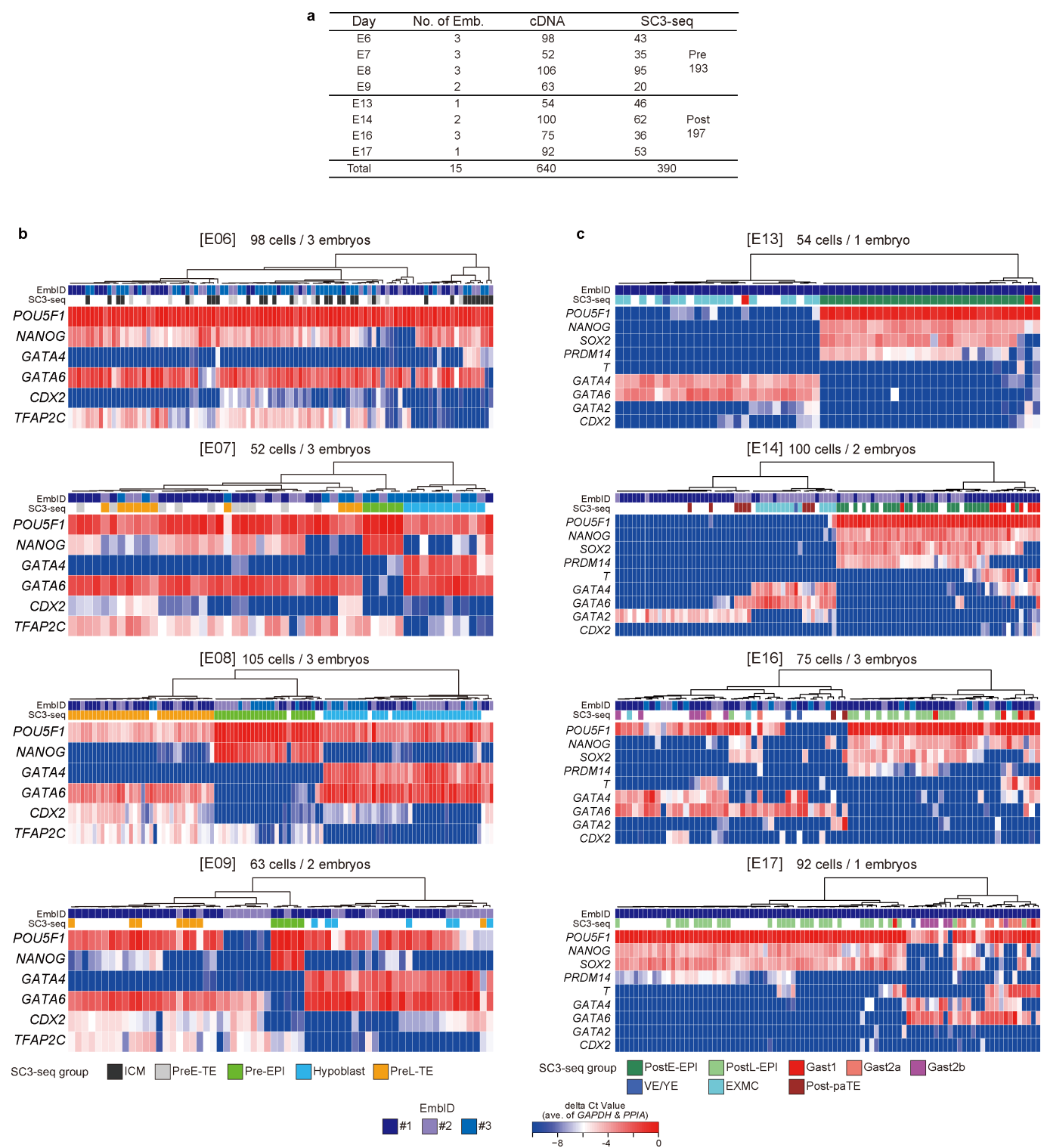
of CDX2/OCT4/GATA4 from E6 to E9 (embryos $n = 7, 6, 12, 12$, respectively). The numbers of cells positive for each marker are indicated. **h**, Expression of GATA6/OCT4 from E6 to E9 (embryos $n = 4, 13, 11, 3$, respectively). The numbers of cells positive for each marker are indicated. **i**, Expression of TFAP2C/OCT4/GATA6 at E8 (embryos $n = 5$). ICM is magnified (right). Arrowheads indicate hypoblast. **j**, Summary of the expression of key markers in monkey, human and mouse pre-implantation embryos. Blastocysts of the three species show grossly similar morphology, but notably, monkey hypoblast extends parietally to cover mural trophectoderm. OCT4 expression appears to be equal in EPI and hypoblast of mouse blastocysts⁵⁶, whereas OCT4 is expressed at a higher level in EPI than in hypoblast and trophectoderm in human and monkey blastocysts^{13,42}. NANOG and GATA4 exhibit a similar expression pattern among the three species^{56–58}. CDX2 shows a similar expression in human and monkeys, but an earlier expression in morula in mice^{13,56}. GATA6 exhibits the most variable expression pattern among the three species: it is expressed only in hypoblast in humans and mice^{57,59}, but is uniformly expressed in hypoblast and trophectoderm in monkeys. Scale bars, 100 μm .



Extended Data Figure 2 | See next page for caption.

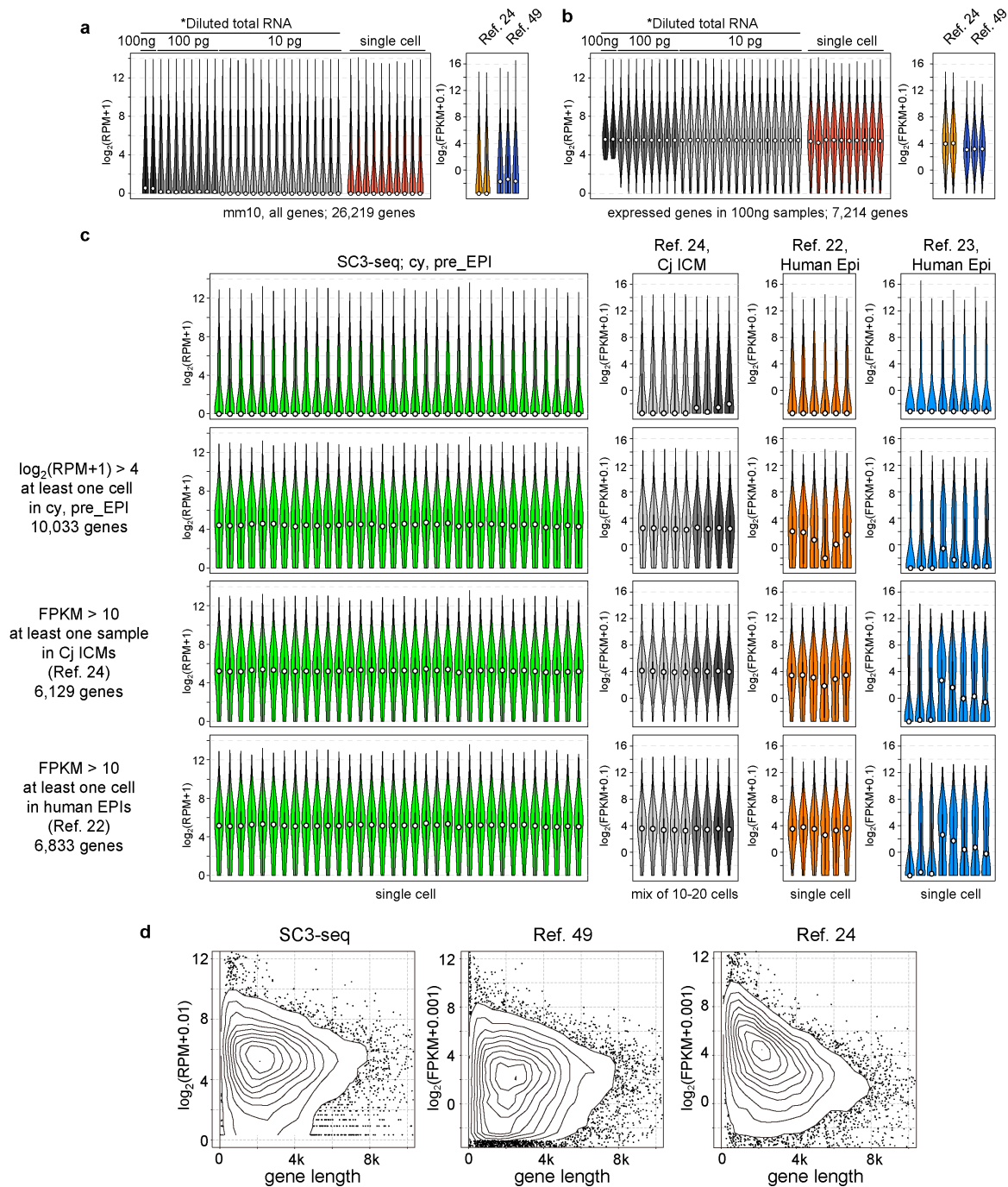
Extended Data Figure 2 | Monkey early post-implantation development. **a**, Ultrasound diagnosis of the recipient uterus for the implantation of transplanted embryos at E14 and E16. Dashed circles indicate the uterus and arrowheads indicate the chorionic cavity. Scale bars, 10 mm. **b**, Implantation (white arrowheads) and pseudo-implantation (black) sites on the recipient endometrium. The image at right is a higher magnification of the area boxed on the left. The implantation site was identified by maternal blood in the trophoblastic lacunae. The pseudo-implantation site was a reacted endometrium to the implantation on the overlying endometrium. Scale bar, 2 mm. **c**, Scheme of monkey early post-implantation development. AM, amnion; CS, connective stalk; CT, cytotrophoblast; EXCM, exocoelomic membrane; EXMC, extra-embryonic mesenchyme; Gast, gastrulating cells;

SYS, secondary yolk sac; VE, visceral endoderm; YE, yolk-sac endoderm; TE, trophoctoderm. **d**, Lower magnification images of Fig. 2b showing whole implantation sites at E14 (left) and E16 (right). PYS/CC, primary yolk sac/chorionic cavity. Scale bars, 500 μ m (left) and 1.0 mm (right). **e**, Expression of OCT4/NANOG/GATA4 and OCT4/T/GATA6 in post-implantation embryos at E16 (embryos $n = 2$). Gastrulating cells positive for T and OCT4 migrated along visceral endoderm. Some cells (yellow arrowhead) showed ingress into visceral endoderm (white arrowhead). Scale bars, 100 μ m. **f**, Expression of OCT4/GATA4 (left) and OCT4/GATA6 (right) in embryos at E14 (embryos $n = 2$). Arrowheads indicate extra-embryonic mesenchyme. Scale bars, 100 μ m.



Extended Data Figure 3 | Expression of key markers in single-cell cDNAs generated from monkey pre- and post-implantation embryos. **a**, Summary of the SC3-seq samples. The numbers of embryos analysed (all of the embryos exhibited normal morphology), of synthesized cDNAs with appropriate quality, and of the cells analysed by SC3-seq are listed. **b**, **c**, qPCR analysis of the expression of key markers in single-cell

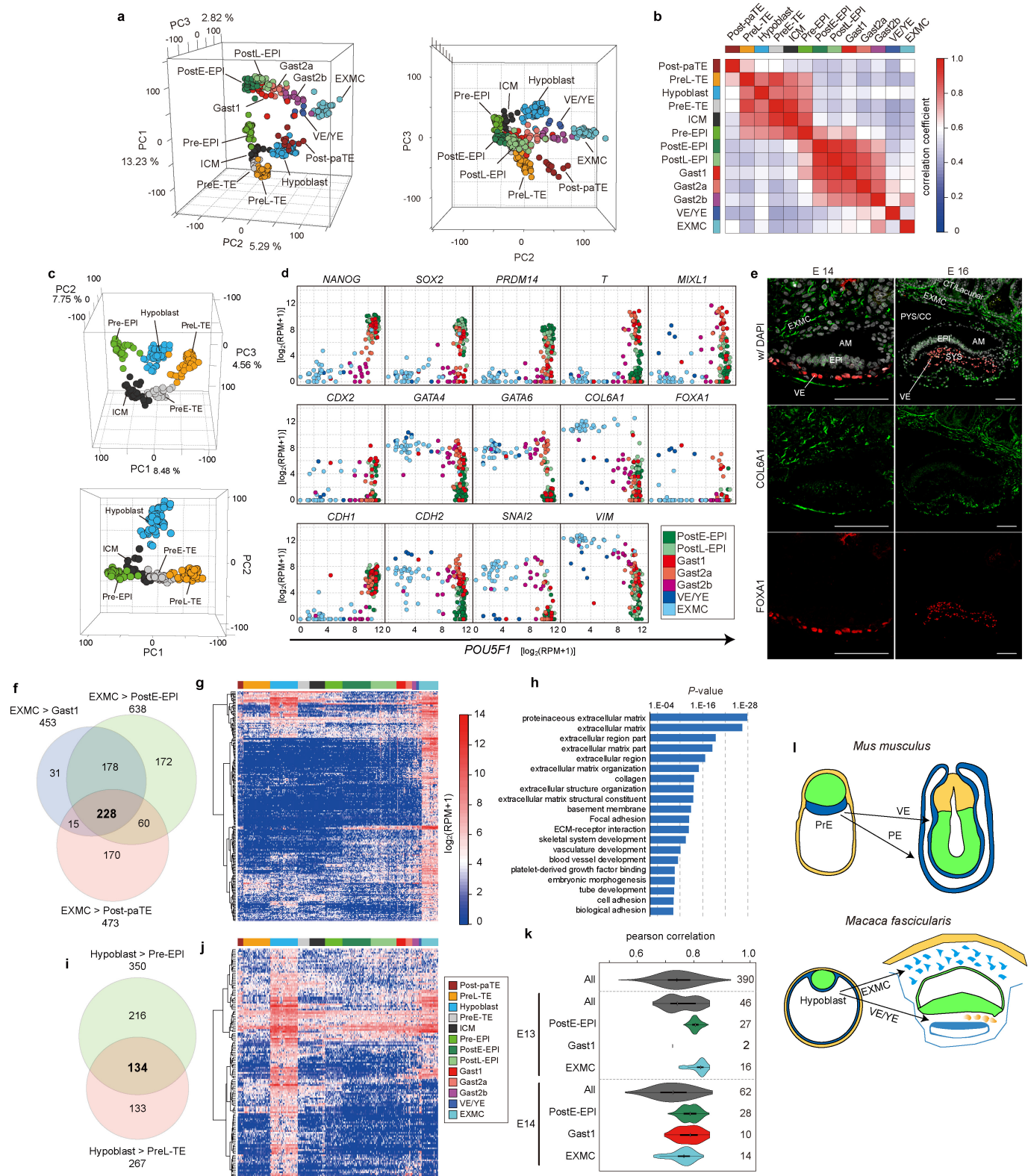
cDNAs generated from pre-implantation (E6, E7, E8, E9) (**b**) and post-implantation (E13, E14, E16, E17) (**c**) embryos. The ΔC_t values from the average C_t values of *GAPDH* and *PPIA* are shown as heat maps and are used for clustering. The identities of the embryos and the samples used for the SC3-seq analyses (annotations are based on Fig. 3a) are indicated. The colour coding is as indicated.



Extended Data Figure 4 | Comparison of the performance of the SC3-seq with that of other RNA-seq and single-cell RNA-seq methods.

a, Distributions of the expression levels of all annotated genes by the SC3-seq (diluted total RNA¹⁹, single cells; in this study) (left) and other methods^{24,49} (right) represented by violin plots. Medians are shown by white circles. For the SC3-seq, the transcript-level ($\log_2(\text{RPM}+1)$) distributions of 100 ng, 100 pg, and 10 pg of total RNA or single cells from 2i+L mESCs (cultured in N2B27 medium supplemented with a cytokine leukaemia inhibitory factor (LIF) and two kinase inhibitors (PD0325901 and CHIR99021))³² as starting materials are shown. For refs 24 and 49, the transcript levels are shown as $\log_2(\text{FPKM}+0.1)$. Transcripts from 20 cells of 2i+L mESCs are amplified in ref. 24, and in ref. 49, 500 ng of polyA RNA from serum/LIF mESC/iPSCs are used for a standard RNA-seq procedure. **b**, Distributions of the expression levels of genes

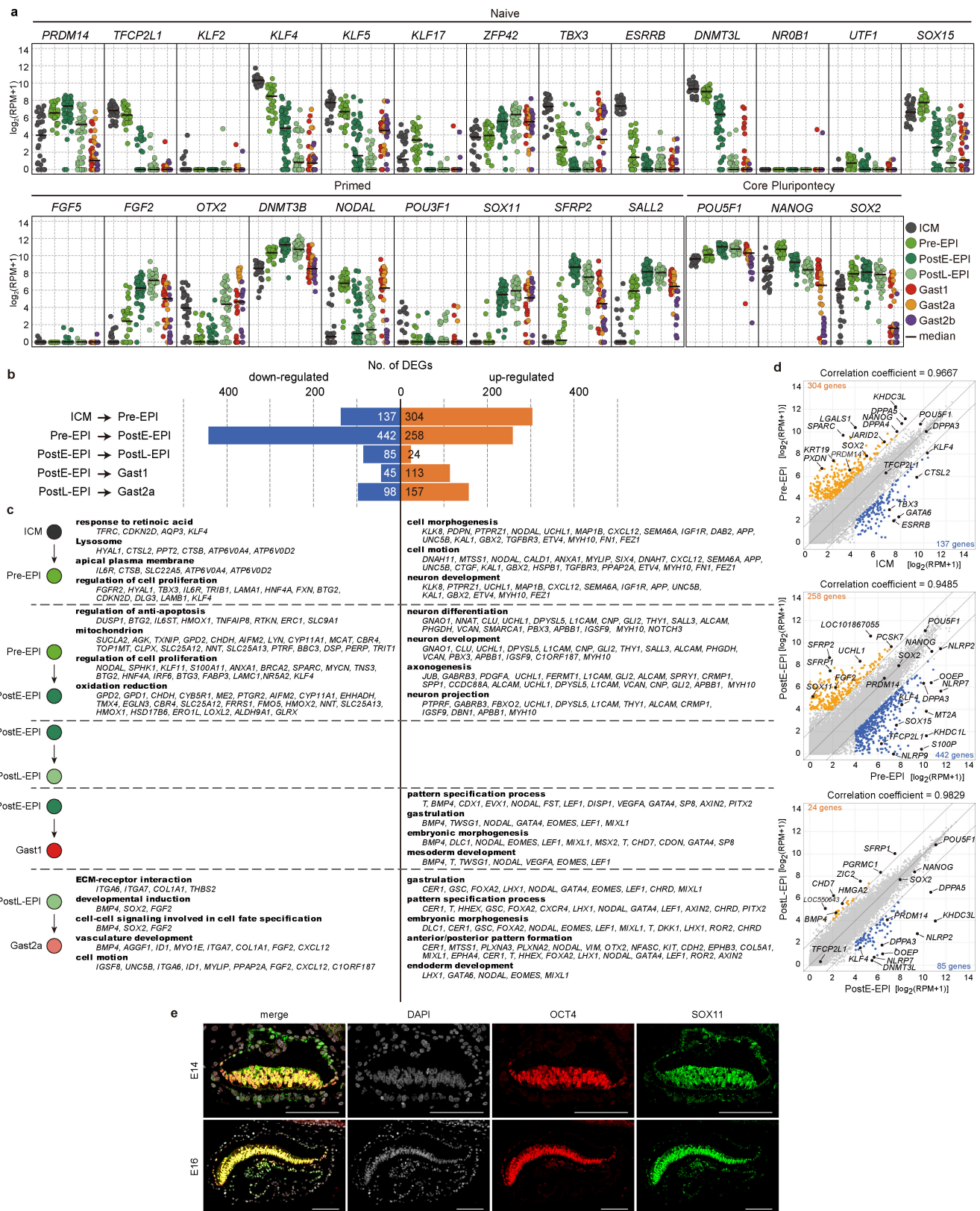
expressed at significant levels ($\log_2(\text{RPM}+1) > 4$ in at least one 100 ng RNA sample) among the same sample set in **a**. **c**, Comparisons of the distributions of the expression levels among corresponding samples in different data sets^{22–24}. The geneset used in the first row consisted of all annotated genes. In the second, third, and fourth rows, the genesets used were genes expressed at significant levels as in **b** in at least one cell in the monkey pre-EPI group, in marmoset ICM samples, and in human EPI samples²² respectively. **d**, Scatter-plot analysis of the correlation between the expression level and the transcript length detected by the SC3-seq (2iLESC_MS68T82)¹⁹, (GSM1119616)⁴⁹, and (ERR637931)²⁴. Note that the method in ref. 24 tends to yield lower estimations of the levels of longer transcripts compared to the SC3-seq and standard RNA-seq.



Extended Data Figure 5 | See next page for caption.

Extended Data Figure 5 | Characterization of extra-embryonic mesenchyme. **a**, PCA of all cells by all expressed genes (390 cells, 18,353 genes). The colour coding is as indicated. **b**, Heat map of correlation coefficients among cells during monkey development. The values were calculated using the averaged expression levels of 6,167 DEGs in each cell type (the genes exhibiting more than fourfold changes among the groups ($\text{FDR} < 0.01$), and the mean of the expression level of at least one group was $>(\log_2(\text{RPM}+1)=4)$). **c**, PCA of cells from pre-implantation embryos by all expressed genes among these groups (193 cells, 15,187 genes). **d**, Scatter-plot comparison of the expression of key genes with that of *POU5F1* in post-implantation cells. **e**, Expression of COL6A1/FOXA1 in embryos at E14 and E16 (embryos $n = 2, 2$, respectively). Scale bars, 100 μm . **f**, Venn diagram showing the overlap among the genes expressed at higher levels in EXMC (>4 -fold) compared to postE-EPI, Gast1,

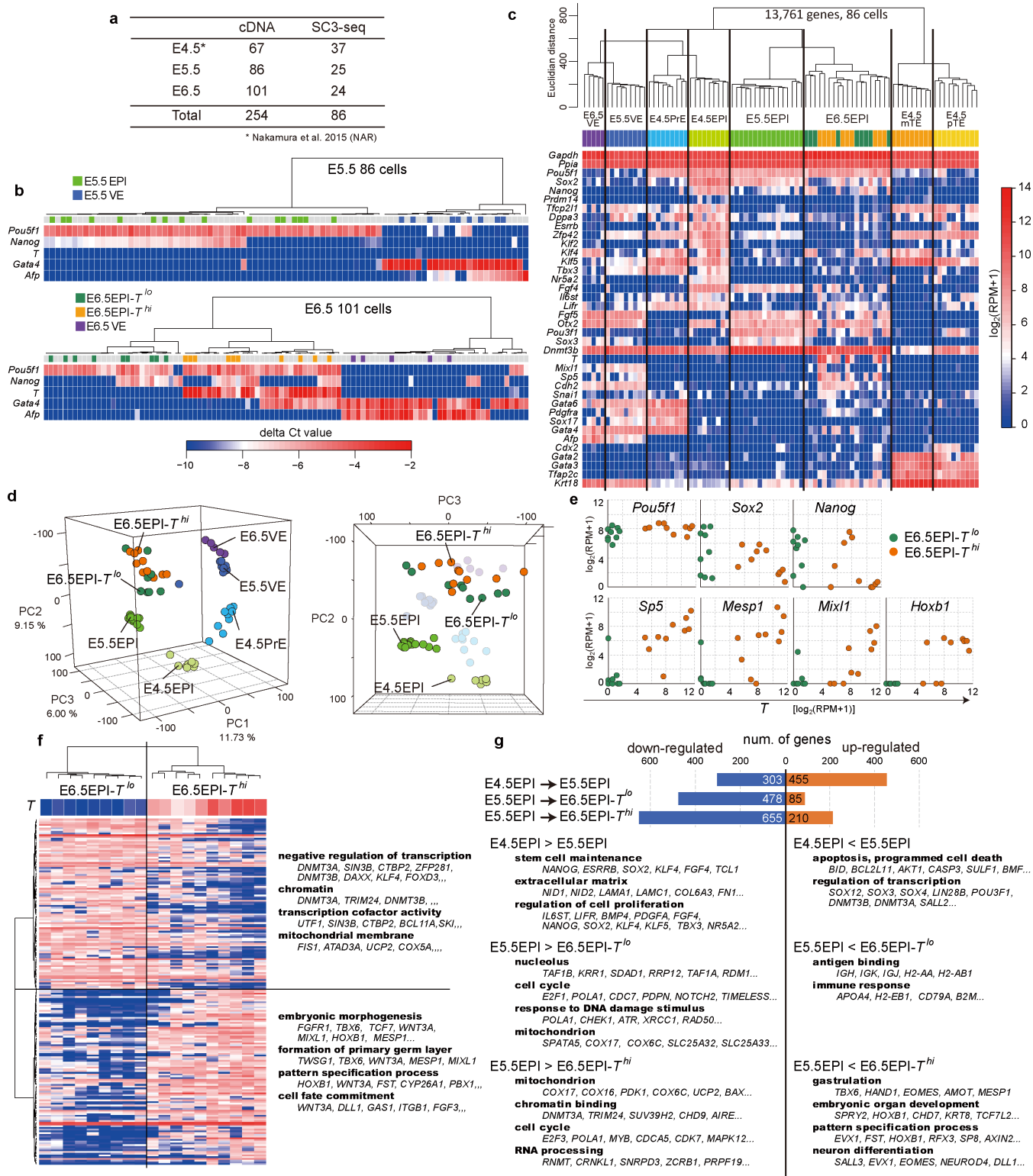
or post-implantation parietal trophectoderm at E13 and E14. **g**, Heat map of the expression of the 228 genes identified, as in **f**, among monkey pre- and post-implantation cell types. **h**, Enrichment of Gene Ontology terms in the 228 genes identified, as in **f**. **i**, Venn diagram showing the overlap between the genes expressed at higher levels in hypoblast (>4 -fold) compared to pre-EPI or preL-TE. **j**, Heat map of the expression of the 134 genes identified, as in **i**, among monkey pre- and post-implantation cell types. **k**, Pearson's correlation coefficients shown by violin plots between all possible pairs of single-cell cDNAs of all cells, postE-EPI, Gast1, and extra-embryonic mesenchyme at E13 and E14. The numbers of cell analysed are indicated on the right. **l**, Schemes for the differentiation of primitive endoderm into visceral or parietal endoderm (VE or PE) in mice (top) or for the differentiation of hypoblast into VE/YE or extra-embryonic mesenchyme in cynomolgus monkeys (bottom).



Extended Data Figure 6 | DEGs during the cyEPI development.

a, Expression of selected genes during EPI development (black bars, median values). **b**, DEGs during the cyEPI development. Orange and blue bars indicate the numbers of up- and downregulated genes, respectively, in the pair-wise comparisons indicated. **c**, Enrichment of Gene Ontology terms and representative genes (all genes are shown in Supplementary Table 2) in DEGs in the pair-wise comparisons indicated. **d**, Scatter-plot comparison of the averaged gene-expression levels between ICM and

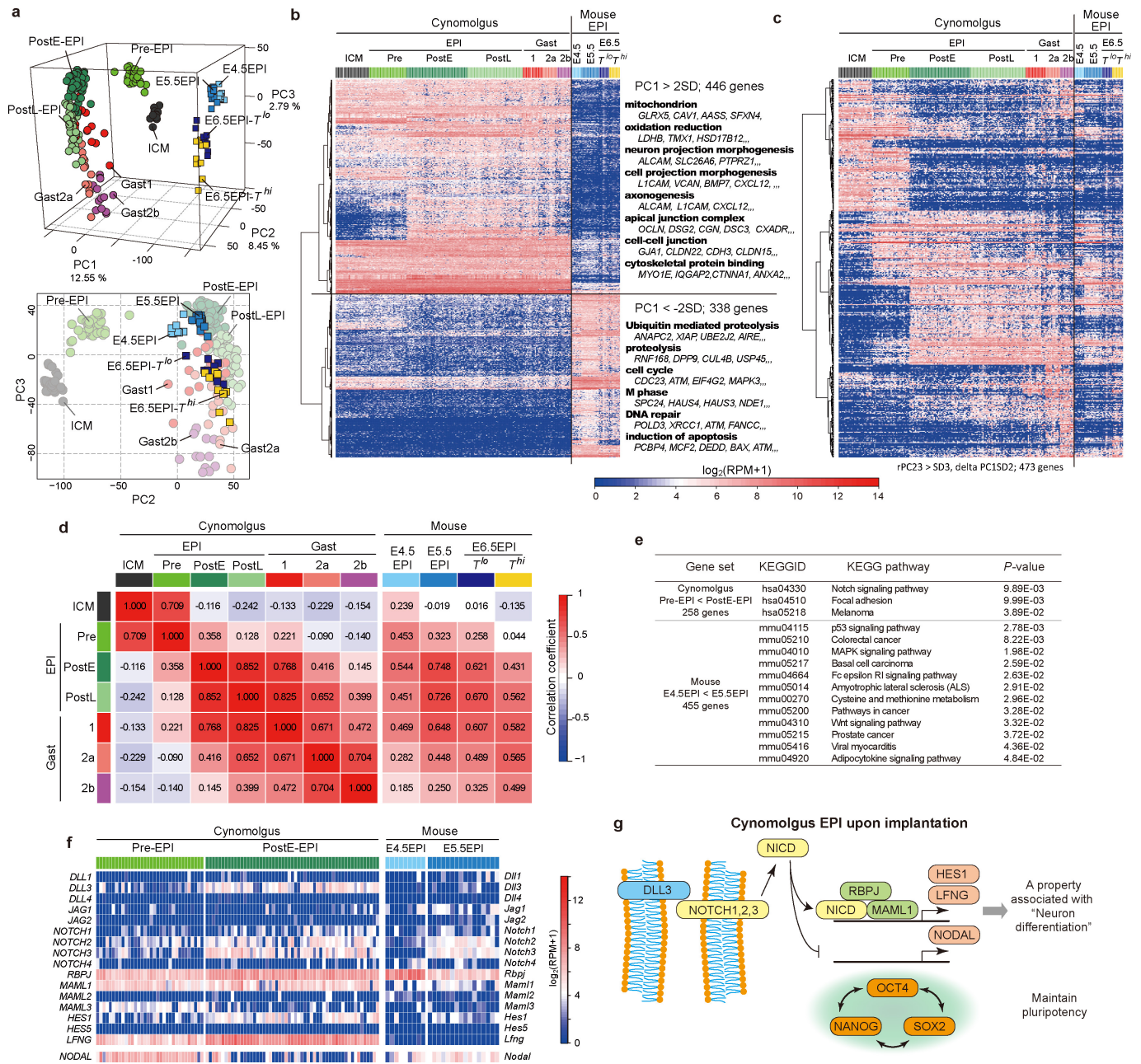
pre-EPI (top), pre-EPI and postE-EPI (middle), and postE-EPI and postL-EPI (bottom). Orange, upregulated; blue, downregulated (>4 -fold difference (flanking diagonal lines), mean $\log_2(\text{RPM}+1) > 4$ in one cell type, FDR < 0.01). Key genes are annotated and the numbers of DEGs are indicated. The correlation coefficient is indicated above the scatter plot. **e**, Expression of SOX11/OCT4 in embryos at E14 (top) and E16 (bottom) (embryos $n = 2, 2$, respectively). Scale bars, 100 μm .



Extended Data Figure 7 | See next page for caption.

Extended Data Figure 7 | SC3-seq analyses of cells in mouse pre- and post-implantation embryos. **a**, Summary of the SC3-seq samples of mouse embryonic cells. The numbers of synthesized cDNAs of appropriate quality and of the cells analysed by SC3-seq are listed. **b**, qPCR analysis of the expression of key markers in single-cell cDNAs from mouse E5.5 (pre-gastrulation) and E6.5 (early/mid-streak stage) embryos. The ΔC_t values from the average C_t values of *Gapdh* and *Arbp* are shown as heat maps and are used for clustering. For E5.5 embryos, cells were picked from EPI and visceral endoderm. For E6.5 embryos, cells were picked from proximal EPI and visceral endoderm. The samples used for the SC3-seq analyses (the annotations are based on **c**) are indicated. **c**, UHC of cells from mouse E4.5 (EPI, primitive endoderm, mural trophoctoderm (mTE), and polar trophoctoderm (pTE)¹⁹), E5.5, and E6.5 embryos by all expressed genes ($\log_2(\text{RPM}+1) > 4$ in at least one sample among 86 cells, 13,761 genes), and heat map of the levels of selected marker genes. Colour bars under the dendrogram indicate the cell types. Orange and green

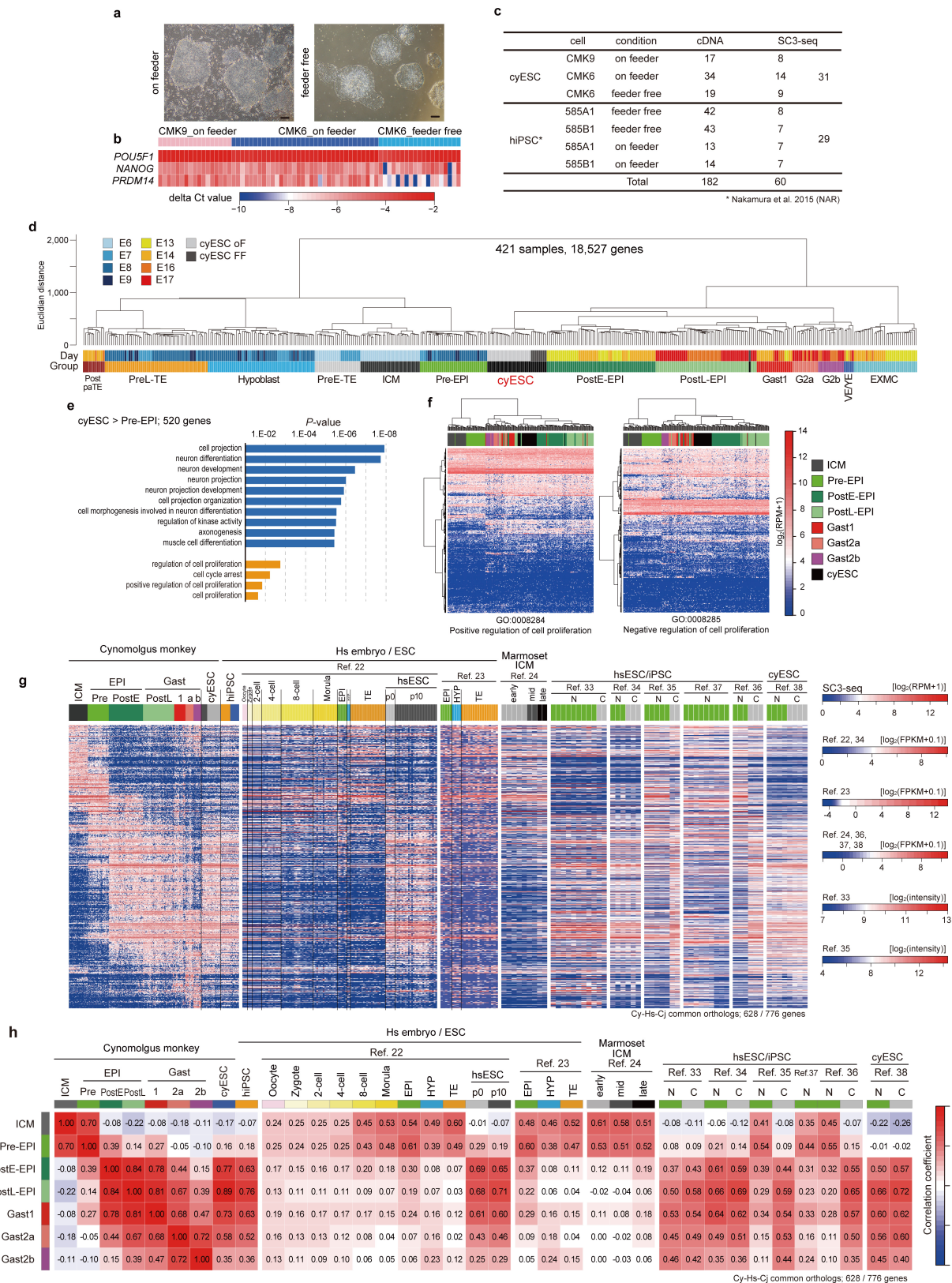
bars in the E6.5 EPI cluster indicate E6.5EPI- T^{lo} (green) and E6.5EPI- T^{hi} (orange) cells, respectively. See **e** for details. **d**, PCA of cells by all expressed genes among the indicated cells (67 cells, 13,761 genes). **e**, Scatter-plot comparison of the expression of key genes for pluripotency or primitive streak formation against that of *T* in E6.5 EPI. Cells were classified by the levels of *T* (orange, E6.5EPI- T^{hi} ; green, E6.5EPI- T^{lo}). **f**, Heat map of the levels of genes with correlation or anti-correlation with *T* in E6.5 EPI. Genes were selected as follows: the levels $> \log_2(\text{RPM}+1) = 6$ in at least one cell, correlation coefficient with *T* > 0.6 (102 genes) or < -0.6 (99 genes). Enrichment of Gene Ontology terms and representative genes are indicated. **g**, DEGs during the mEPI development. Top, Orange and blue bars indicate the numbers of up- and downregulated genes, respectively, in the pair-wise comparisons indicated. Bottom, Enrichment of Gene Ontology terms and representative genes in DEGs in the pair-wise comparisons indicated.



Extended Data Figure 8 | Comparison of monkey and mEPI development.

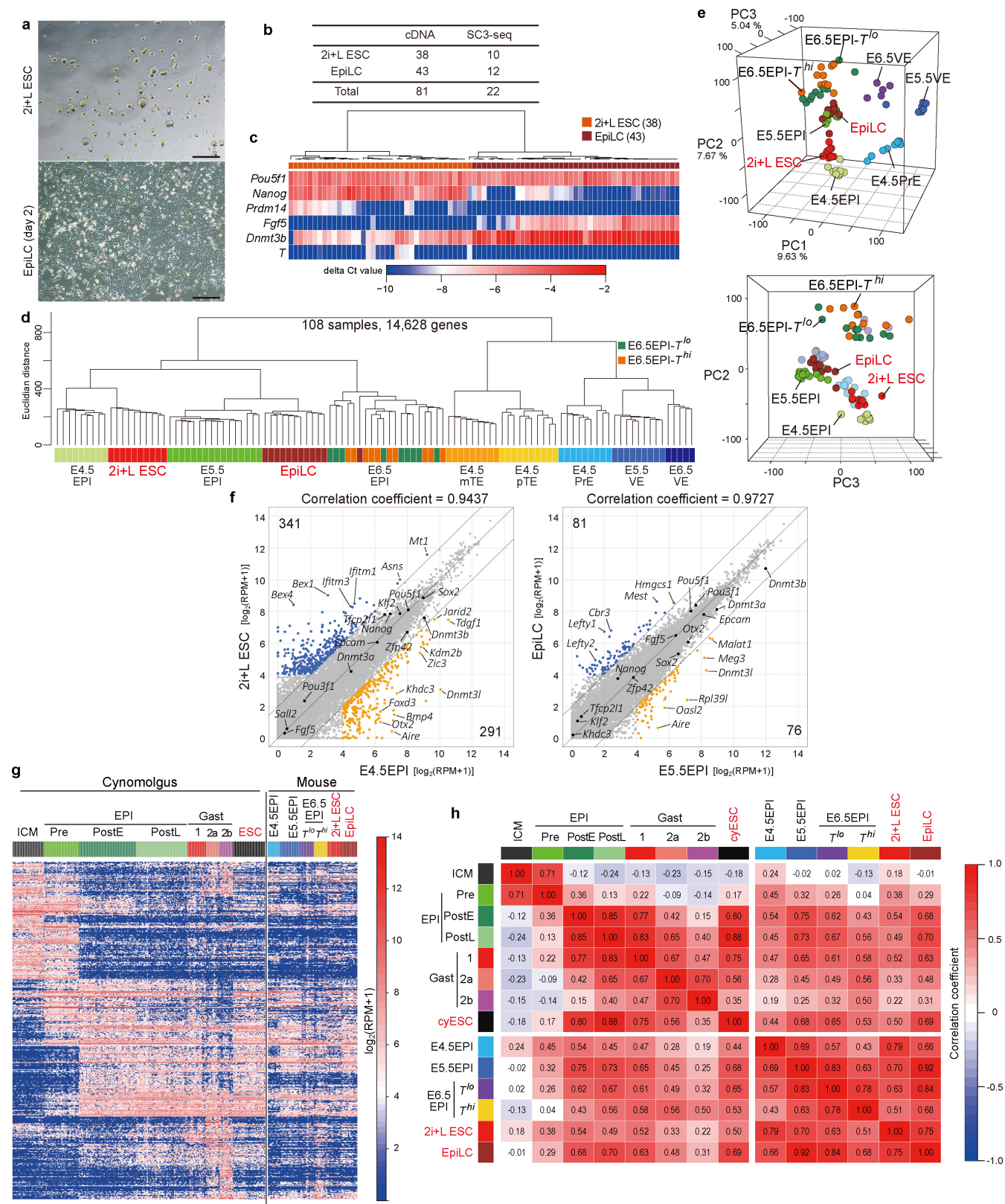
a, PCA of cells during monkey (circles, colour-coded as in Fig. 3a) and mouse (squares) EPI development. Orthologues among humans, cynomolgus monkeys and mice were annotated (15,220 genes), and 13,473 genes expressed among these cells (monkey, 213 cells; mouse, 44 cells) were used for PCA. **b**, Heat map of the expression of 784 genes that contributed highly to the PC1 axis (>2 s.d. of PC1: cyEPI or mEPI genes (PCA as in **a**)). The genes are ordered by UHC, and representative cyEPI or mEPI genes and their key Gene Ontology enrichments are shown. **c**, Heat map of the levels of monkey and mouse common EPI genes (473 genes) (defined as: radius of PC2 and 3 > 3 s.d. and of PC1: -2 s.d.

< PC1 < 2 s.d.) during monkey and mEPI development. **d**, Heat map of correlation coefficients among cells during cyEPI and mEPI development. The values were calculated using the averaged expression level of monkey and mouse common EPI genes (473 genes (**c**, Supplementary Table 2)). **e**, Signalling pathways enriched in genes upregulated during the pre-EPI to postE-EPI transition (top, 258 genes) or during the E4.5–E5.5 mEPI transition (bottom, 455 genes) by the KEGG pathway analysis. **f**, Heat map of the expression changes of key genes in the NOTCH pathway and NODAL/Nodal during upon implantation. **g**, A proposed pathway operating in monkey post-EPI, which acquires a property for 'neuron differentiation'. NICD, NOTCH intracellular domain.



Extended Data Figure 9 | Correlations between hPSCs and cyPSCs and cells during cyEPI development. **a**, Morphology of cyESCs (CMK6) cultured with (left) or without (right) feeders. Scale bars, 200 μ m. **b**, qPCR analysis of the expression of key markers in single-cell cDNAs generated from cyESCs (CMK6 and CMK9) cultured with or without feeders. The ΔC_t values from the average C_t value of *GAPDH* and *PPIA* are shown as heat maps. **c**, Summary of the SC3-seq samples of cyESCs and hiPSCs¹⁹. The numbers of synthesized cDNAs of appropriate quality and of the cells analysed by SC3-seq are listed. **d**, UHC with all expressed genes (421 cells, 18,527 genes). Note that one cyESC is clustered with postL-EPI.

e, Enrichment of Gene Ontology terms in genes upregulated in cyESC against pre-EPI (520 genes). **f**, UHC with expression of genes for 'positive' (left, 351 genes of Gene Ontology: 0008284) or negative (right, 391 genes of Gene Ontology: 0008284) regulation of cell proliferation. **g**, Heat map of the expression of cyEPI ontogenic genes among the indicated cells, including those reported by others^{22–24,33–38}. The genes in common for all platforms were used (628/776 genes). N, 'naive'; C, conventional. **h**, Heat map of the correlation coefficients among cells as in **g**. Correlation coefficients were calculated using the averaged expression levels of genes in **g**.



Extended Data Figure 10 | See next page for caption.

Extended Data Figure 10 | Correlations among mPSCs and cells during cyEPI and mEPI development. **a**, Morphology of mESCs with 2i+L and day 2 epiblast-like cells (EpiLC) induced from the mESCs³². Scale bars, 200 μ m. **b**, Summary of the SC3-seq samples of mESCs and epiblast-like cells. The numbers of synthesized cDNAs of appropriate quality and of the cells analysed by SC3-seq are listed. **c**, qPCR analysis of the expression of key markers in single-cell cDNAs generated from 2i+L mESCs and epiblast-like cells. The ΔC_t values from the C_t values of *Arbp* are shown as heat maps and are used for clustering. **d**, UHC of cells from E4.5, E5.5, and E6.5 embryos, 2i+L mESCs, and epiblast-like cells with all expressed genes (108 cells, 14,628 genes). Colour bars under the dendrogram indicate the cell types. **e**, PCA of mEPI, primitive endoderm/visceral

endoderm, 2i+L mESCs, and epiblast-like cells by all expressed genes among these cells (89 cells, 14,341 genes). **f**, Scatter-plot comparisons of the averaged gene-expression levels between E4.5 EPI and 2i+L mESCs (left), and between E5.5 mEPI and epiblast-like cells (right). Key genes are annotated and the numbers of DEGs are indicated. The correlation coefficient is indicated above the scatter plots. **g**, Heat map of the expression of monkey and mouse common EPI genes (473 genes, as in Extended Data Fig. 8b) in cells during cyEPI and mEPI development and in cyPSCs and mPSCs. **h**, Heat map of the correlation coefficients among cells as in **g**. Correlation coefficients were calculated using the averaged expression levels of genes in **g**.

Tumour hypoxia causes DNA hypermethylation by reducing TET activity

Bernard Thienpont^{1,2*}, Jessica Steinbacher^{3*}, Hui Zhao^{1,2*}, Flora D'Anna^{1,2*}, Anna Kuchnio^{1,4}, Athanasios Ploumakis⁵, Bart Ghesquière¹, Laurien Van Dyck^{1,2}, Bram Boeckx^{1,2}, Luc Schoonjans^{1,4}, Els Hermans⁶, Frederic Amant⁶, Vessela N. Kristensen^{7,8}, Kian Peng Koh⁹, Massimiliano Mazzone^{1,10}, Mathew L. Coleman⁵, Thomas Carell³, Peter Carmeliet^{1,4} & Diether Lambrechts^{1,2}

Hypermethylation of the promoters of tumour suppressor genes represses transcription of these genes, conferring growth advantages to cancer cells. How these changes arise is poorly understood. Here we show that the activity of oxygen-dependent ten-eleven translocation (TET) enzymes is reduced by tumour hypoxia in human and mouse cells. TET enzymes catalyse DNA demethylation through 5-methylcytosine oxidation. This reduction in activity occurs independently of hypoxia-associated alterations in TET expression, proliferation, metabolism, hypoxia-inducible factor activity or reactive oxygen species, and depends directly on oxygen shortage. Hypoxia-induced loss of TET activity increases hypermethylation at gene promoters *in vitro*. In patients, tumour suppressor gene promoters are markedly more methylated in hypoxic tumour tissue, independent of proliferation, stromal cell infiltration and tumour characteristics. Our data suggest that up to half of hypermethylation events are due to hypoxia, with these events conferring a selective advantage. Accordingly, increased hypoxia in mouse breast tumours increases hypermethylation, while restoration of tumour oxygenation abrogates this effect. Tumour hypoxia therefore acts as a novel regulator of DNA methylation.

Although the mutagenic processes underlying oncogenesis are well studied, tumours are known to be not only genetically but also epigenetically distinct from their tissue of origin. The most extensively documented examples of oncogenic epigenetic changes are those to DNA methylation, but the underlying mechanisms are poorly understood¹.

In tumours, changes in DNA methylation involve both global hypomethylation and the local hypermethylation of CpG-rich gene promoters¹. Hypermethylation frequently affects tumour suppressor genes (TSGs), downregulating their expression and thus contributing to oncogenesis. It remains unclear how methylation changes arise, but an instructive model suggests that genetic changes are a prerequisite for methylation changes²; *BRAF* mutations, for instance, lead to hypermethylation in colorectal tumours³. This is problematic as, while pervasive, hypermethylation of TSGs can only be explained by somatic mutations in a fraction of tumours. Notably, extensive hypermethylation can be seen in ependymomas completely devoid of somatic mutations⁴.

In contrast to DNA methylation mechanisms, those of demethylation have remained elusive until recently, when TET methylcytosine dioxygenases (TET1, TET2 and TET3) were shown to oxidize 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC)⁵. 5-Hydroxymethylcytosine and its further-oxidized derivatives are subsequently replaced with an unmodified cytosine by base-excision repair to achieve demethylation⁶. Reduced 5mC oxidation due to decreased TET activity thus increases levels of DNA methylation. Mutations suppressing TET activity are often found in myeloid leukaemia and glioblastoma^{6–9}, but less frequently in other cancer types. By contrast, 5hmC loss is pervasive in tumours and even proposed as a cancer hallmark¹⁰. As with hypermethylation, somatic mutations explain the loss of 5hmC in only a fraction of tumours and it remains unclear which other factors trigger this loss².

Notably, like hypoxia-inducible factor (HIF)-prolyl-hydroxylase domain proteins (PHDs), TET enzymes are Fe²⁺- and α -ketoglutarate-dependent dioxygenases¹¹. PHDs are oxygen-sensitive, acting as oxygen sensors. Under normoxic conditions, they hydroxylate the HIF transcription factors, targeting them for proteasomal degradation, whereas under hypoxia they do not, leading to HIF stabilization and hypoxia response activation¹². Expanding tumours continuously become disconnected from their vascular supply, resulting in vicious cycles of hypoxia, HIF activation and tumour vessel formation¹³. Consequently, hypoxia pervades in solid tumours. Oxygen levels range from 5% to anoxia and around one-third of tumour areas contain less than 0.5% oxygen¹⁴. Although DNA hypermethylation and hypoxia are well-recognized cancer hallmarks, the effect of hypoxia on TET hydroxylase activity and subsequent DNA de-methylation has not been assessed. We therefore set out to investigate whether a hypoxic micro-environment decreases TET hydroxylase activity in tumours, leading to an accumulation of 5mC and acquisition of hypermethylation.

Effect of hypoxia on DNA hydroxymethylation

To assess whether hypoxia affects TET activity, we exposed ten human and five murine cell lines with detectable 5hmC levels to 21% O₂ (normoxic) or 0.5% O₂ (hypoxic, commonly observed in tumours¹⁴) for 24 h. Hypoxia induction was verified and DNA was extracted and profiled for nucleotide composition using liquid chromatography–mass spectrometry (LC–MS). We observed 5hmC loss in eleven cell lines, including eight cancer cell lines (Fig. 1a). However, this did not translate into global 5mC increases (Extended Data Fig. 1), presumably because 5mC is more abundant and is not targeted by TETs at many sites¹⁵. The effect of hypoxia was concentration- and time-dependent:

¹Vesalius Research Center, VIB, 3000 Leuven, Belgium. ²Laboratory of Translational Genetics, Department of Oncology, KU Leuven, 3000 Leuven, Belgium. ³Center for Integrative Protein Science, Department für Chemie und Pharmazie, Ludwig-Maximilians-Universität, 81377 München, Germany. ⁴Laboratory of Angiogenesis and Vascular Metabolism, Department of Oncology, KU Leuven, 3000 Leuven, Belgium. ⁵Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham B15 2TT, UK. ⁶Gynecologic Oncology, University Hospitals Leuven, Department of Oncology, KU Leuven, 3000 Leuven, Belgium. ⁷Department of Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, N-0310 Oslo, Norway. ⁸Department of Clinical Molecular Biology (EpiGen), Akershus University Hospital and Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Postboks 1171, Blindern 0318 Oslo, Norway. ⁹Department of Development and Regeneration, and Stem Cell Institute Leuven, KU Leuven, 3000 Leuven, Belgium. ¹⁰Laboratory of Molecular Oncology and Angiogenesis, Department of Oncology, KU Leuven, 3000 Leuven, Belgium.

*These authors contributed equally to this work.

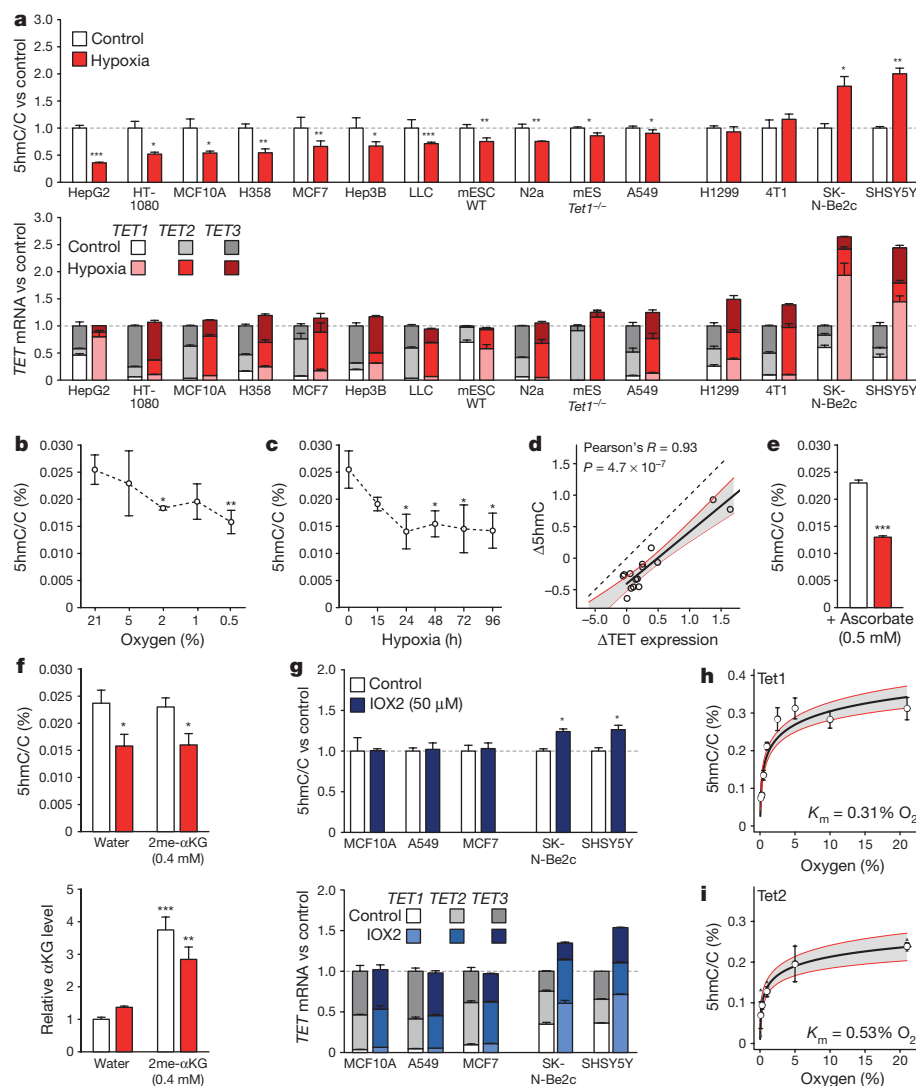


Figure 1 | Effect of hypoxia on 5hmC *in vitro*. **a**, Levels of 5hmC (top), and overall *TET* expression (bottom) in cell lines grown for 24 h under 21% or 0.5% O₂. RNA expression is expressed relative to the combined estimated level of all 3 *TET* paralogues under 21% O₂. **b**, **c**, 5hmC/C levels in MCF7 cells exposed to different O₂ levels for 24 h (**b**), or 0.5% O₂ for indicated times (**c**). **d**, Correlation of changes in overall *TET* expression and 5hmC upon hypoxia. Each circle represents a cell line, the full line the correlation. **e**, **f**, Levels of 5hmC (**e**) and α -ketoglutarate (α KG) (**f**) in

MCF7 cells grown with ascorbate (**e**), water or dimethyl- α -ketoglutarate (2me- α KG) (**f**) under 21% (white) or 0.5% (red) O₂. α -ketoglutarate changes are relative to matching water controls. **g**, As in **a**, but for cells exposed to IOX2. **h**, **i**, Michaelis-Menten curve of Tet1 (**h**) and Tet2 (**i**, $n = 3$) for O₂. K_m denotes Michaelis constant. Error bars denote s.e.m., grey areas: 95% confidence interval, $n = 5$ replicates for **a**–**h**, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ by *t*-test (**b**, **c**, **e**) or analysis of variance (ANOVA) with post-hoc Tukey HSD test (**f**).

a dose–response revealed loss of 5hmC at oxygen levels at-or-below 2%, and a 20% and 40% reduction, respectively, after 15 h and >24 h of hypoxia (Fig. 1b, c). Loss of 5hmC was not due to increased 5hmC oxidation to 5fC¹⁶, as hypoxia also decreased 5fC levels in embryonic stem (ES) cells (Extended Data Fig. 1).

In some cell lines, levels of 5hmC failed to decrease under hypoxia. 5hmC levels were unaffected in cell lines H1299 and 4T1, and even increased in SHSY5Y and SK-N-Be2c neuroblastoma cells, as reported previously¹⁷ (Fig. 1a). When profiling *TET* expression, neuroblastoma cells displayed potent hypoxia-dependent induction of *TET1* and *TET2*. Cell lines H1299 and 4T1 exhibited intermediate increases in expression levels, whereas all other cell lines showed no, or modest, increases of some *TET* paralogues (Fig. 1a). *Tet* gene expression changes were confirmed at the protein level in mouse cell lines, and HIF1 β -chromatin immune precipitation followed by sequencing (ChIP-seq) further confirmed that HIF binds near the promoters of upregulated *Tet* genes, but not near those that are unaltered (Extended Data Fig. 2a, b), in keeping with the cell-type specificity of the hypoxia response¹². Notably, no cell line showed

decreased *Tet* expression, indicating that 5hmC loss is not due to reduced *Tet* expression.

Since hypoxia affects *TET* paralogue expression differently in different cell lines we correlated hypoxia-associated changes in overall *TET* expression (the combined abundances of *TET1*, *TET2* and *TET3*) with changes in 5hmC levels. Hypoxia reduced 5hmC levels by an average of 44% ($P = 0.0097$) in each cell line (Fig. 1d), independently of *TET* expression changes. Nevertheless, changes in *TET* expression also affected 5hmC levels. This was confirmed by short interfering RNA (siRNA) knockdown of *TET2*, which constitutes around 60% of all *TET* expression in MCF7 cells. This reduced 5hmC levels by around 60% (Extended Data Fig. 2c). Similarly, *Tet1* knockout mouse ES cells (*Tet1*^{−/−}) displayed lower 5hmC levels than wild-type ES cells, in which *Tet1* is the predominantly expressed *Tet* paralogue under both normoxic and hypoxic conditions (Fig. 1a, Extended Data Fig. 2d).

Post-hypoxic 5hmC levels therefore appear to be determined by altered oxygen availability and by changes in *TET* abundance. This explains why cell lines that do not upregulate *TET* expression in response to hypoxia display 5hmC loss, whereas cell lines that strongly

upregulate *TET* compensate for this, resulting in equal or increased 5hmC levels.

Hypoxia directly affects DNA hydroxymethylation

Aside from gene expression, *TET* activity is affected by a variety of cellular processes, including changes in levels of reactive oxygen species (ROS), Krebs cycle metabolites and proliferation^{7,11,17,18}. Since such changes might also occur secondary to hypoxia, we investigated whether they underlie 5hmC reductions in hypoxia.

First, we assessed whether ROS could affect *TET*s in the nucleus through inactivation of Fe^{2+} in their catalytic domain. Although hypoxia increased overall ROS levels, no increase in nuclear ROS was detected either by a nucleus-specific ROS probe or through 8-oxoguanine (8-oxoG) quantification (Extended Data Fig. 3a–f). Moreover, ascorbate supplementation to counteract ROS increases¹⁹ failed to rescue 5hmC loss (Fig. 1e).

Second, because changes in metabolites such as succinate and fumarate compete with *TET* for its cofactor α -ketoglutarate⁷, we investigated whether this was relevant. The concentration of these metabolites, however, was not increased in hypoxic MCF10A or embryonic stem (ES) cells, and only 3–4-fold in MCF7 cells (Extended Data Fig. 3g–i). Levels of the onco-metabolite 2-hydroxyglutarate were also increased in hypoxic MCF7 and MCF10A cells, but were still only around 5–10% of α KG (Extended Data Fig. 3h, j), and therefore unlikely to affect *TET* activity, as affinity of these competing metabolites for hydroxylases is lower or similar to α KG^{7,20}. Culturing MCF7 cells in glutamine-free medium to decrease the concentration of these metabolites did not alter 5hmC levels (Extended Data Fig. 3k). Similarly, exogenous addition of cell-permeable α KG under hypoxia to counteract putative competing metabolites did not rescue the 5hmC loss (Fig. 1f). This therefore precluded metabolite competition from causing hypoxia-associated 5hmC loss.

Third, increases in cell proliferation have also been linked to 5hmC loss²¹. However, cell growth was unaffected or decreased upon exposure to hypoxia in all cell lines tested, indicating that increased proliferation does not underlie 5hmC reduction (Extended Data Fig. 3l).

Fourth, to exclude any potential cellular changes caused by HIF activation, we pharmacologically activated the hypoxia response program by exposing five cell lines grown in atmospheric conditions to IOX2, a small molecule inhibitor with high specificity for PHDs²² (Extended Data Fig. 3m). Cell lines not characterized by hypoxia-induced *TET*-expression changes (MCF10A, A549 and MCF7) showed no change in 5hmC levels under IOX2, while those characterized by *TET* upregulation (SK-N-BE2c and SHSY5Y) showed an increase in 5hmC (Fig. 1g). Thus, after IOX2 exposure, changes in 5hmC levels mirrored changes in *TET* transcription. We also prepared nuclear protein extracts from MCF7 cells grown under hypoxic and atmospheric conditions, and then compared their 5mC oxidative capacities at the same oxygen tension *in vitro*. These, however, were identical (Extended Data Fig. 3n). Loss of 5hmC was therefore not due to activation of the hypoxia response program.

Finally, we assessed the effect of varying oxygen concentrations on the activity of recombinant purified Tet1 or Tet2, by measuring conversion of 5mC to 5hmC on double-stranded genomic DNA. We observed a dose-dependent reduction in 5hmC production with decreasing concentrations. Importantly, under the hypoxic conditions applied in this study (0.5% O_2), Tet1 and Tet2 activity were reduced by $45\% \pm 7$ and $52\% \pm 8$ (mean \pm s.e.m., $P = 0.01$; Fig. 1h, i).

Together, these data demonstrate that decreased oxygen availability directly diminishes the oxidative activity of *TET*s, independently of changes in HIF activity, competing metabolites, proliferation, nuclear ROS or *TET* expression.

Loci with differential DNA hydroxymethylation

To analyse where in the genome hypoxia reduces 5hmC. DNA from hypoxic and normoxic MCF7 cells was immunoprecipitated using

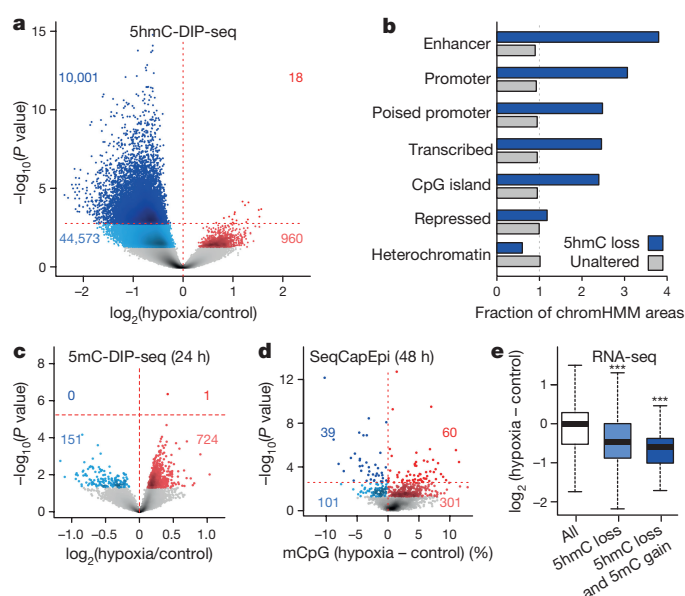


Figure 2 | Genomic profiles of 5(h)mC in MCF7 following hypoxia.

a, Changes in 5hmC at 290,382 peaks detected using 5hmC-DIP-seq. Peaks gaining (red) and losing (blue) 5hmC are highlighted at $P < 0.05$ and 5% FDR adjustment (lighter and darker). **b**, Observed/expected fraction of 5hmC peaks overlapping with chromHMM chromatin states either exhibiting hypoxia-associated 5hmC loss ($n = 10,001$, blue) or not ($n = 280,381$, grey). **c**, **d**, Changes in 5mC after 24 h (**c**) or 48 h (**d**) of 0.5% O_2 , assessed by 5mC-DIP-seq at 10,001 hypohydroxymethylated peaks upon hypoxia (**c**) or by BS-seq at 1,894 regions capture-selected using SeqCapEpi (**d**; see Methods). **e**, Expression changes of genes in hypohydroxymethylated, and both hypohydroxymethylated and hypermethylated peaks. Plots depict 3 (**a**, **e**), 4 (**c**) or 5 (**d**) replicates, *** $P < 0.001$ by negative binomial generalized linear models (**a**, **c**), Fisher's exact test (**d**) or *t*-test (**e**).

antibodies targeting 5mC or 5hmC and subjected to high-throughput sequencing (DIP-seq). We detected 290,382 sites enriched for 5hmC. After hypoxia, 10,001 of the peaks generated for each site exhibited a decrease in 5hmC (5% false discovery rate (FDR)) and only 18 exhibited an increase, thereby confirming global 5hmC loss (Fig. 2a, Supplementary Table 1). Genomic annotation of these peaks using chromHMM²³ revealed they were predominantly found at gene promoters, with some at enhancers and actively transcribed regions, in line with known *TET*-binding sites¹⁵ (Fig. 2b). For example, 5hmC was decreased near transcription start sites of TSGs *NSD1*, *FOXA1* and *CDKN2A* (Extended Data Fig. 4). Analysis of 5mC-DIP signals at these 10,001 regions highlighted that, in 724 out of 875 altered regions, the 5mC content was significantly increased ($P < 0.05$), although only one of these sites survived a 5% FDR correction (Fig. 2c, Supplementary Table 2). Increases in 5mC were thus more subtle than decreases observed for 5hmC.

Several days may be required for 5hmC changes to translate into 5mC changes¹⁹. We therefore cultured cells for 48 h (rather than 24 h) under hypoxia, and used targeted bisulfite-sequencing (BS-seq) to obtain base-resolution quantitation of 5mC at around 85 Mb of promoters and enhancers. Using this approach, we could assess increases in 5mC for 1,894 of the 10,001 regions displaying 5hmC loss. As observed after 5mC-DIP-seq, 301 out of 402 altered sites displayed increased methylation ($P < 0.05$). Similarly, 60 out of 99 altered sites were increased with 5% FDR correction ($P = 2.8 \times 10^{-3}$; Fig. 2d, Supplementary Table 3). ChromHMM annotation revealed that these 60 sites were predominantly in gene promoters and enhancers. To assess the effect of hypermethylation on gene expression, we performed RNA sequencing (RNA-seq) on hypoxic MCF7 cells. Genes depleted in 5hmC and with increased 5mC showed significantly decreased expression upon hypoxia (Fig. 2e; $P = 2.5 \times 10^{-42}$ and 7.4×10^{-4} , respectively, for 3,660

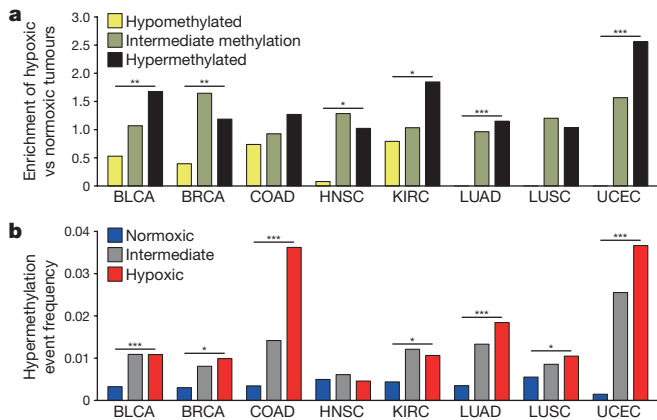


Figure 3 | Effect of hypoxia on hypermethylation in TCGA.

a, Observed and expected number of hypoxic versus normoxic tumours in 3 methylation clusters for 1,000 CpGs hypermethylated in tumour versus normal tissue. **b**, Percentage of hypermethylation events in promoters of frequently hypermethylation genes. $n = 3,141$ tumours, $*P < 0.05$, $**P < 0.01$, $***P < 0.001$ by Cochran–Armitage (**a**), generalized linear model per tumour type corrected for co-variants (Supplementary Table 8) (**b**). BLCA, bladder carcinoma; BRCA, breast carcinoma; COAD, colorectal adenocarcinoma; HNSC, head and neck squamous cell carcinoma; KIRC, kidney renal clear cell carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; UCEC, uterine corpus endometrial carcinoma.

genes with 5hmC loss and 55 genes with both 5hmC loss and 5mC gain; Supplementary Table 4). Reduced TET activity therefore leads to an accumulation of 5mC, decreasing the expression of associated genes.

Hypermethylation events in hypoxic tumours

We next assessed whether 5hmC loss and concomitant 5mC gain also occur *in vivo*. We focused on gene promoters as they are more frequently affected upon hypoxia, and directly linked to gene expression. As cancer cells go through multiple rounds of sustained hypoxia¹⁴, we proposed that there would be an increase in 5mC, as it would provide a selective advantage for cancer cells, similar to somatic mutations. We therefore assessed 5hmC levels in three patient-derived tumour xenografts, in which we marked hypoxic areas with pimonidazole (Extended Data Fig. 5a). Immunofluorescence analysis revealed decreased 5hmC in hypoxic areas, linking tumour hypoxia to 5hmC loss *in vivo*.

To assess whether hypoxia-associated hypermethylation contributes to the oncogenic process, we analysed tumours profiled in the pan-cancer study of The Cancer Genome Atlas (TCGA)²⁴. We selected 8 solid tumour types (3,141 tumours) for which both DNA methylation (450k array) and gene expression (RNA-seq) data were available for >100 samples, and classified each as hypoxic, normoxic or intermediate using an established gene signature²⁵ (Extended Data Fig. 5b). Next, we analysed tumour-associated DNA hypermethylation in each tumour type by performing unsupervised clustering of 1,000 CpGs that displayed the strongest hypermethylation in tumour versus normal tissue (Extended Data Fig. 5c). In the first three clusters (displaying low, intermediate and high average hypermethylation), we analysed the enrichment of hypoxic tumours. For all eight tumour types, hypoxic tumours predominated in the hypermethylated cluster and normoxic tumours in the hypomethylated cluster (Fig. 3a; $P = 2 \times 10^{-4}$), suggesting that hypoxia leads to increased methylation in tumours.

Whereas the prior analysis identified uniform increases in methylation based on average changes, it poorly captured exceptional increases in hypermethylation known to occur in a subset of tumours^{1,26}. We therefore also modelled tumour hypermethylation by annotating increases in CpG methylation at gene promoters using a stringent threshold (Bonferroni-corrected $P < 0.05$) as hypermethylation events. In each tumour type, the promoters of 187 ± 38 out of 29,649 genes frequently displayed hypermethylation events (Supplementary Table 5).

Notably, hypoxic tumours had on average 4.8-fold more hypermethylation events in these genes than normoxic tumours (Fig. 3b; $P = 4.1 \times 10^{-13}$). These events were functional, reducing gene expression in tumours carrying these hypermethylation events (Extended Data Fig. 5d). They primarily affected promoters with high or intermediate CpG content, in line with TET target preference (Extended Data Fig. 5e)¹⁵. Furthermore, they were not restricted to a small subset: $77\% \pm 6.5$, $49\% \pm 9.3$ or $39\% \pm 9.1$ of hypoxic tumours were affected by ≥ 1 , ≥ 10 or ≥ 20 hypermethylation events, respectively. Considering hypermethylation frequency in normoxic tumours as baseline, up to 48% of hypermethylation events were hypoxia-related.

As hypermethylation can also be genetically encoded, mutations in some genes correlated positively with hypermethylation (for example, *IDH1*, *TET1*, *TET3* and *BRAF*; Supplementary Table 6). Importantly, hypoxia predicted hypermethylation independent of mutation status ($P = 6.1 \times 10^{-12}$). Mutations inhibiting TET activity were infrequent (approximately 1.8% of tumours), indicating that hypermethylation is not genetically encoded in most tumours. *TET*-mutant tumours were also not more hypoxic, suggesting that hypoxia induces hypermethylation, and not vice versa (Extended Data Fig. 5f). Hypoxia-associated hypermethylation events occurred independently of other tumour characteristics such as tumour cell percentage, immune cell infiltration, tumour size, proliferation or metastasis ($P = 4 \times 10^{-13}$), and were significant in seven out of eight tumour types (Supplementary Tables 7, 8). In line with an earlier report²¹, high proliferation was the only other variable significantly predicting hypermethylation ($P = 5.3 \times 10^{-10}$), although only in four of eight tumour types (Extended Data Fig. 5g, h). Using multiple regression, we estimated the contribution of tumour characteristics to hypermethylation variance. On the basis of partial correlation coefficients, proliferation predicted $12.1\% \pm 4.1$, and hypoxia $33.3\% \pm 5.7$, of hypermethylation events explained by the model (Extended Data Fig. 5i).

Given the increase in hypermethylation events in hypoxic tumours, we next selected genes with more hypermethylation events in hypoxic versus normoxic tumours (5% FDR). This revealed 263 ± 94 genes per tumour type, with $9.0\% \pm 1.6$ being shared between any two types (Supplementary Table 9). Ontology analysis of hypermethylated genes revealed that they had biological processes in common such as cell cycle arrest, DNA repair and apoptosis. Hypermethylation was also observed in genes involved in suppressing glycolysis, angiogenesis and metastasis, consistent with tumour hypoxia inducing these processes (Extended Data Fig. 6a–c).

Reduced TET activity underlies hypermethylation

We used three strategies to confirm the role of TET activity in hypoxia-associated hypermethylation. First, we correlated *TET* expression with hypermethylation events, correcting for hypoxia and proliferation. *TET2* and *TET3* expression inversely correlated with hypermethylation ($P = 0.046$ and 0.0028 , Extended Data Fig. 7a), as did hypoxia and proliferation ($P < 1.2 \times 10^{-13}$ for both). Similar to our *in vitro* observations, this implicates reduced TET activity in hypermethylation.

Second, we assessed the overlap of hypermethylation events induced by hypoxia and *IDH1*^{R132} mutations⁸ in 63 glioblastomas. Among wild-type *IDH1* glioblastomas, hypermethylation frequency was 3.4-fold higher in hypoxic tumours (Fig. 4a, Extended Data Fig. 7b). As expected, *IDH1*^{R132} tumours were hypermethylated, albeit 3.9-fold more so than hypoxic tumours (Fig. 4a). This indicates that TET enzymes, fully inactivated in *IDH*-mutant tumours⁹, were only partially inactivated in hypoxia, similar to our *in vitro* observations. Of 228 genes frequently hypermethylated in glioblastomas, those in the hypoxic and *IDH*-mutant subgroups displayed a 58% overlap ($P < 10^{-16}$; Fig. 4b) and reduced expression (Extended Data Fig. 7c), indicating that loss of TET activity affects the same genes, regardless of the underlying trigger.

Finally, to link hypoxia-associated hypermethylation to 5hmC loss, we profiled 24 non-small-cell lung tumours for 5mC and 5hmC using 450k arrays (Extended Data Fig. 7d). This revealed a generalized loss

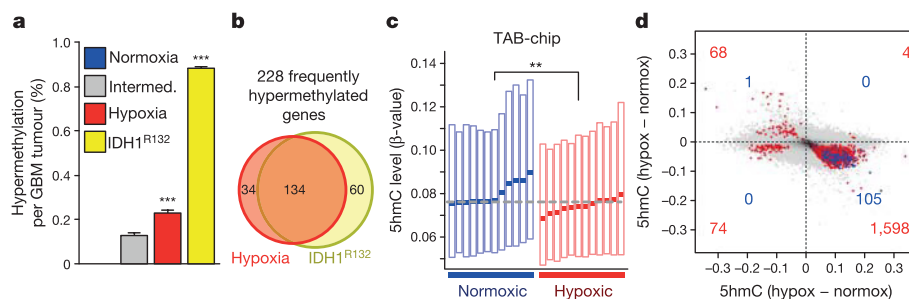


Figure 4 | Effect of hypoxia on TET activity in human tumours.

a, Hypermethylation in 19 normoxic (blue), 21 intermediate (grey), 17 hypoxic (red) and 6 IDH1^{R132}-mutated (yellow) glioblastomas. **b**, Overlap between genes hypermethylated in hypoxic versus IDH1^{R132}-mutated glioblastomas. **c**, 5hmC measured across 485,000 CpGs in 12

normoxic versus 12 hypoxic non-small-cell lung tumours. **d**, Changes in 5(h)mC for unaltered CpGs (grey), and CpGs altered in both 5mC and 5hmC (25% FDR, blue; $P < 0.01$, red). *** $P < 0.001$ by Fisher's exact (**a**), ** $P < 0.01$ by t -test (**c**).

of 5hmC in hypoxic tumours ($-7.1\% \pm 1.1$; $P = 3.7 \times 10^{-3}$; Fig. 4c). Individual probes also mostly displayed 5hmC loss and 5mC gain in hypoxic tumours (96.7% and 65.4% of probes altered, respectively, $P < 0.01$; Supplementary Table 10). Of all probes displaying 5mC gain, most (87%) also displayed 5hmC loss, and of probes altered both in 5hmC and 5mC ($P < 0.01$), 92% showed 5hmC loss and 5mC gain (Fig. 4d; $P < 10^{-16}$). This directly implicates hypoxia-induced loss of 5hmC in the hypermethylation of hypoxic tumours.

Rescue of hypoxia-induced hypermethylation

To manipulate tumour oxygenation and confirm its effect on hypermethylation, we used mice expressing the polyomavirus middle T-antigen under the mouse mammary tumour virus promoter (MMTV-PyMT). These mice spontaneously develop breast tumours, with hypoxic areas emerging from 7 weeks onwards, encompassing approximately 20% of the tumour at 16 weeks²⁷. Hypoxic areas in these tumours were also depleted in 5hmC (Fig. 5a, b).

We monitored hypermethylation changes by targeted BS-seq of TSG promoters commonly inactivated in cancer²⁸. Hypoxic human breast tumours displayed a specific increase in hypermethylation at these TSG promoters, whereas no effect was observed for oncogenes (Extended Data Fig. 8a). In line with the age-associated increase in tumour hypoxia²⁷, hypermethylation events also increased markedly with age (and tumour size), but not in normal mammary glands (Extended Data Fig. 8b–d). Importantly, >95% of cells in these tumours were PyMT-positive, whereas cell proliferation and immune cell infiltration were comparable between hypoxic and normoxic areas (Extended Data Fig. 8e–g). Hypermethylation changes are therefore unlikely to be a result of changes in proliferation or cellular heterogeneity.

To test whether reduced tumour oxygenation increases hypermethylation, 9-week-old MMTV-PyMT mice were hydrodynamically injected with a soluble-Flk1 (sFlk1)-expressing plasmid. After 3 weeks, this caused tumour vessel pruning and hypoxia (Extended Data Fig. 9a–d). Shallow whole-genome sequencing for 5hmC (TET-assisted bisulfite sequencing; TAB-seq) revealed a global loss of 5hmC after sFlk1 overexpression ($-12.4\% \pm 3.5$, $P = 0.040$), occurring predominantly at gene-dense regions and affecting the entire gene (Fig. 5c, Extended Data Fig. 9e), consistent with previously described 5hmC distributions¹⁵. Moreover, targeted BS-seq revealed an exacerbated hypermethylation phenotype after sFlk1 overexpression at 12 weeks in TSGs but not oncogenes (10 out of 15 TSGs contained ≥ 1 hypermethylation event; $P = 0.010$, Fig. 5d, Extended Data Fig. 9f). Tumour growth and the expression of proliferation markers, *Tet* paralogues and the immune cell marker CD45 were unaffected by sFlk1 overexpression, indicating that hypermethylation occurs independently (Extended Data Fig. 9g–j).

To rescue this effect, we normalized the tumour vasculature by intercrossing a heterozygous *Phd2* (also known as *Egln1*) loss-of-function allele with the PyMT transgene. This significantly reduced tumour hypoxia at 16 weeks²⁷ (Extended Data Fig. 9k). TAB-seq revealed a 5hmC gain ($+25.3\% \pm 4.7$, $P = 0.0098$) occurring primarily

at gene-dense regions and affecting the entire gene (Fig. 5c, Extended Data Fig. 9l). Notably, BS-seq revealed that, although 8 out of 15 TSGs displayed ≥ 1 hypermethylation event in *Phd2*^{+/+} tumours, no hypermethylation was observed in *Phd2*^{+/−} tumours ($P = 2.6 \times 10^{-7}$, Fig. 5e). Again, oncogenes were unaffected (Extended Data Fig. 9m). Effects were independent of *Phd2* haploinsufficiency in tumour cells, as similar effects were observed in PyMT mice having endothelial-cell-specific *Phd2* haploinsufficiency²⁷ (Extended Data Fig. 9n, o). As in the sFlk1

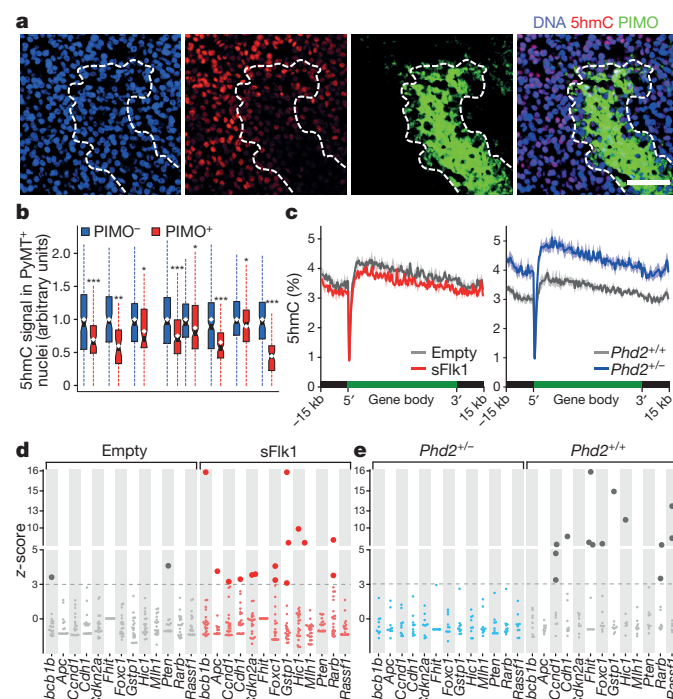


Figure 5 | Effect of vessel pruning and normalization on 5hmC and TSG hypermethylation.

a, **b**, Immunofluorescence of breast tumours in transgenic (MMTV-PyMT) mice. **a**, Representative image. Scale bar, 50 μ m. **b**, Box plot of 5hmC signal in >150 PyMT-positive nuclei from eight tumours, stratified for pimonidazole (PIMO) (yes/no) and normalized to PIMO-negative nuclei. **c**, 5hmC levels \pm s.e.m. across a metagene in tumours of 12-week-old mice receiving empty or sFlk1-overexpressing plasmid (left, $n = 3$), or 16-week-old mice with the indicated genotype (right, $n = 3$ for *Phd2*^{+/+}; $n = 4$ for *Phd2*^{+/−}). **d**, **e**, Hypermethylation in tumours developing in 12-week-old mice receiving empty ($n = 19$) or sFlk1-overexpressing plasmid ($n = 24$) 3 weeks earlier (**d**), and in tumours developing in 16-week-old *Phd2*^{+/−} ($n = 10$) and *Phd2*^{+/+} ($n = 9$) mice (**e**). Plotted are z-scores of hypermethylation, relative to normoxic tumours (empty and *Phd2*^{+/−} for **d** and **e**). Dotted line: 5% FDR, darker dots: significant hypermethylation. *Brca1* and *Timp3*: not shown (no hypermethylation event detected). Hypermethylated genes on average had 5.8% (**d**) and 4.7% (**e**) more methylation. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ by t -test.

model, increasing tumour oxygenation by *Phd2* haploinsufficiency did not affect tumour growth or the expression of proliferation markers, *Tet* paralogs or CD45 (Extended Data Fig. 9p–u).

Discussion

We show here that tumour hypoxia directly reduces TET activity, causing a 5hmC decrease predominantly at gene promoters and enhancers. Concomitantly, 5mC increases at these sites and, as with certain genetic mutations, provides a substrate for oncogenic selection *in vivo*²⁶. Since hypoxia prevails in tumours, 5mC changes in TSG promoters are frequent, rendering hypoxic tumours hypermethylated at these sites. Hypermethylation events in tumours have long been suspected to occur through selection of random DNA methylation variants²⁹. However, the identification of genetically encoded hypermethylation challenged this stochastic model². By demonstrating that hypoxia drives hypermethylation, we show that genetically-encoded and tumour-microenvironment-driven models of epimutagenesis co-exist. However, since hypoxia is pervasive, the mechanism described here is relevant for most solid tumours. We found that up to 48% of hypermethylation events were hypoxia-related and effects were replicated in all tumour types investigated, independent of mutation- and proliferation-induced hypermethylation. Modest hypoxia (2–5% O₂) did not affect TET activity, indicating that TET enzymes are not physiological oxygen sensors (unlike PHDs) in line with previous reports³⁰. TET activity only becomes limiting under the pathophysiological oxygen concentrations found in tumours¹⁴. Analogous to somatic *TET* haploinsufficiency, this partial reduction in TET activity contributes to oncogenesis. Our findings also suggest intriguing avenues of investigation into other ischaemia-related pathologies.

Our model provides a mechanism for the association between hypoxia and maladaptive oncogenic processes. Genes affected by hypermethylation were not only involved in cell-cycle arrest, DNA repair and apoptosis, but also glycolysis, metastasis and angiogenesis. High doses of angiogenesis inhibitors stimulate metastatic spreading in mouse cancer models (at least in specific settings)³¹, and tumour hypoxia is considered a driver of this behaviour. The mechanism by which hypermethylation accumulates under hypoxia may underlie these escape mechanisms. By contrast, low levels of angiogenic inhibition can induce tumour vessel normalization, and improve oxygenation³². Our observations in normalized PyMT tumours suggest that the therapeutic benefits of vessel normalization such as decreased metastatic burden²⁷, might occur by inhibiting hypoxia-associated hypermethylation. Countering hypermethylation by inhibiting DNA methylation or by normalizing tumour blood supply may therefore prove to be therapeutically beneficial.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 June 2015; accepted 5 July 2016.

Published online 17 August 2016.

1. Esteller, M. Epigenetics in cancer. *N. Engl. J. Med.* **358**, 1148–1159 (2008).
2. Struhl, K. Is DNA methylation of tumour suppressor genes epigenetic? *eLife* **3**, e02475 (2014).
3. Weisenberger, D. J. et al. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with *BRAF* mutation in colorectal cancer. *Nat. Genet.* **38**, 787–793 (2006).
4. Mack, S. C. et al. Epigenomic alterations define lethal CIMP-positive endophenotypes of infancy. *Nature* **506**, 445–450 (2014).
5. Tahliliani, M. et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).
6. Shen, L. et al. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153**, 692–706 (2013).
7. Xiao, M. et al. Inhibition of α -KG-dependent histone and DNA demethylases by fumarate and succinate that are accumulated in mutations of FH and SDH tumor suppressors. *Genes Dev.* **26**, 1326–1338 (2012).
8. Figueroa, M. E. et al. Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer Cell* **18**, 553–567 (2010).
9. Xu, W. et al. Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α -ketoglutarate-dependent dioxygenases. *Cancer Cell* **19**, 17–30 (2011).

10. Yang, H. et al. Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. *Oncogene* **32**, 663–669 (2013).
11. Ploumaki, A. & Coleman, M. L. OH, the places you'll go! Hydroxylation, gene expression, and cancer. *Mol. Cell* **58**, 729–741 (2015).
12. Schofield, C. J. & Ratcliffe, P. J. Oxygen sensing by HIF hydroxylases. *Nat. Rev. Mol. Cell Biol.* **5**, 343–354 (2004).
13. Hanahan, D. & Folkman, J. Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. *Cell* **86**, 353–364 (1996).
14. Vaupel, P., Höckel, M. & Mayer, A. Detection and characterization of tumor hypoxia using pO₂ histography. *Antioxid. Redox Signal.* **9**, 1221–1235 (2007).
15. Williams, K. et al. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343–348 (2011).
16. Ito, S. et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300–1303 (2011).
17. Mariani, C. J. et al. TET1-mediated hydroxymethylation facilitates hypoxic gene induction in neuroblastoma. *Cell Reports* **7**, 1343–1352 (2014).
18. Zhao, B. et al. Redox-active quinones induces genome-wide DNA methylation changes by an iron-mediated and Tet-dependent mechanism. *Nucleic Acids Res.* **42**, 1593–1605 (2014).
19. Blaschke, K. et al. Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells. *Nature* **500**, 222–226 (2013).
20. Koivunen, P. et al. Transformation by the (R)-enantiomer of 2-hydroxyglutarate linked to EGLN activation. *Nature* **483**, 484–488 (2012).
21. Bachman, M. et al. 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat. Chem.* **6**, 1049–1055 (2014).
22. Chowdhury, R. et al. Selective small molecule probes for the hypoxia inducible factor (HIF) prolyl hydroxylases. *ACS Chem. Biol.* **8**, 1488–1496 (2013).
23. Taberlay, P. C., Statham, A. L., Kelly, T. K., Clark, S. J. & Jones, P. A. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. *Genome Res.* **24**, 1421–1432 (2014).
24. Weinstein, J. N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
25. Buffa, F. M., Harris, A. L., West, C. M. & Miller, C. J. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *Br. J. Cancer* **102**, 428–435 (2010).
26. Feinberg, A. P. & Irizarry, R. A. Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc. Natl Acad. Sci. USA* **107** (Suppl. 1), 1757–1764 (2010).
27. Kuchnio, A. et al. The cancer cell oxygen sensor PHD2 promotes metastasis via activation of cancer-associated fibroblasts. *Cell Reports* **12**, 992–1005 (2015).
28. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
29. Landan, G. et al. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat. Genet.* **44**, 1207–1214 (2012).
30. Laukka, T. et al. Fumarate and succinate regulate expression of hypoxia-inducible genes via TET enzymes. *J. Biol. Chem.* **291**, 4256–4265 (2016).
31. Paez-Ribes, M. et al. Antiangiogenic therapy elicits malignant progression of tumors to increased local invasion and distant metastasis. *Cancer Cell* **15**, 220–231, (2009).
32. Heist, R. S. et al. Improved tumor vascularization after anti-VEGF therapy with carboplatin and nab-paclitaxel associates with survival in lung cancer. *Proc. Natl Acad. Sci. USA* **112**, 1547–1552 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank G. Peuteman, T. Van Brussel, J. Serneels and K. Kurz for assistance, C. Chang for NucPE1, G.-L. Xu for Tet-triple knockout ES cells. H.Z. and B.T. hold FWO-F postdoctoral fellowships. This work was supported by funding from the ERC (CHAMELEON 617595 to D.L.; EU-ERC269073 to P.C.; CHAMELEO 334420 to B.T.), from the FWO-F (G065615N, G070615N) to D.L., from the IUAP (P7/03) and the Flemish Government (Methusalem) to P.C., and from the DFG (EXC114 (CIPSM), grants CA275/8-5, GRK2062/1 and SPP1784) to T.C.

Author Contributions B.T. and D.L. conceived and supervised the project, designed experiments and wrote the manuscript. B.T. and F.D.A. performed *in vitro* experiments and analysed data, helped by L.V.D.; M.L.C. and A.P. analysed Tet Michaelis–Menten kinetics; animal models were provided by E.H., F.A. (xenografts), M.M. (sFlk1), A.K. and P.C. (*Phd2*^{+/−}); V.N.K. contributed ideas, L.S. and K.P.K. provided reagents; J.S. quantified nucleotides by LC–MS, supervised by T.C.; B.G. quantified metabolites. H.Z. analysed TCGA tumours; B.T., H.Z. and B.B. performed bioinformatics and statistics.

Author Information Microarray and sequencing data are available at the Gene Expression Omnibus under accession GSE71403. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.T. (bernard.thienpont@vib-kuleuven.be) or D.L. (diether.lambrechts@vib-kuleuven.be).

Reviewer Information *Nature* thanks R. S. Johnson, M. Rehli and Y. Xiong for their contribution to the peer review of this work.

METHODS

Materials. All materials were molecular biology grade. Unless noted otherwise, all were from Sigma.

Analysis of global 5mC and 5hmC levels in cultured cells. MCF7, MCF10A, A549, H1299, SHSY5Y, Hep G2, Hep 3B2, HT-1080, NCI-H358, LLC, Neuro-2a, 4T1 and SK-N-Be2c cell lines were obtained from the American Type Culture Collection and their identity was not further authenticated. These cell lines are not listed in the database of commonly misidentified cell lines maintained by ICLAC. LLC, Neuro-2a, 4T1, Hep G2, HT-1080, Hep 3B2, MCF7 and A549 cells were cultured at 37 °C in DMEM with 10% fetal bovine serum (FBS), 5 ml of 100 U ml⁻¹ penicillin-streptomycin (Life Technologies) and 5 ml of L-glutamine 200 mM. NCI-H358, H1299 and SK-N-Be2c cell lines were cultured at 37 °C in RPMI 1640 Medium with 10% FBS 1% penicillin-streptomycin and 1% L-glutamine. MCF10A cells were cultured at 37 °C in DMEM/F-12 supplemented with 5% horse serum (Life Technologies), 20 ng ml⁻¹ human epidermal growth factor (Prepote), 0.5 µg ml⁻¹ hydrocortisone, 100 ng ml⁻¹ cholera toxin, 10 µg ml⁻¹ insulin, and 100 U ml⁻¹ penicillin-streptomycin. The SHSY5Y cell line was cultured at 37 °C in DMEM/F-12 supplemented with 10% FBS, 2% penicillin-streptomycin and 1% non-essential amino acids (MEM). Mouse J1 ES cells were cultured feeder-free in fibroblast-conditioned medium. Cell cultures were confirmed to be mycoplasma-free every month.

Cell line treatment conditions. Control cell cultures were grown at atmospheric oxygen concentrations (21%) with 5% CO₂. To render cultures hypoxic, they were incubated in an atmosphere of 0.5% O₂, 5% CO₂ and 94.5% N₂. Where indicated, IOX2 (50 µM), ascorbate (0.5 mM, a dose known to support TET activity¹⁹) or dimethyl- α -ketoglutarate (0.5 mM) was added to fresh culture medium, using an equal volume of the carrier (DMSO) as a control for IOX2. Cells were plated at a density tailored to reach 80–95% confluence at the end of the treatment. Fresh medium was added to the cells just before hypoxia exposure. For glutamine-free culture experiments, dialysed FBS was added to glutamine-free DMEM, and supplemented with glutamine (4 mM) for the control. Mouse J1 ES cells and *Tet1*-gene-trap ES cells were cultured feeder-free in fibroblast-conditioned medium.

DNA extraction. After exposure to the aforementioned stimuli, cultured cells were washed on ice with ice-cold PBS with deferoxamin (PBS-DFO, 200 µM), detached using cell scrapers and collected by centrifugation (400g, 4 °C). Nucleic acids were subsequently extracted using the Wizard Genomic DNA Purification kit (Promega) according to instructions. All buffers were supplemented with DFO (200 µM) and DNA was dissolved in 80 µl PBS-DFO with RNase A (200 U, NEB) and incubated for 10 min at 37 °C. After proteinase K addition (200 units) and incubation for 30 min at 56 °C, DNA was purified using the QIAquick blood and tissue kit (all buffers supplemented with DFO). It was eluted in 100 µl of a 10 mM Tris, 1 mM EDTA solution (pH 8) and stored at –8 °C until further processing.

LC-ESI-MS/MS of DNA to measure 5mC, 5hmC and 8-oxoG levels. To measure the cytosine, 5mC, 5hmC and 8-oxoG content of the DNA samples, three technical replicates were run for each sample. More specifically, 0.5–2 µg DNA in 25 µl H₂O were digested in an aqueous solution (7.5 µl) of 480 µM ZnSO₄, containing 42 U nuclease S1, 5 U Antarctic phosphatase, and specific amounts of labelled internal standards were added and the mixture was incubated at 37 °C for 3 h in a Thermomixer comfort (Eppendorf). After addition of 7.5 µl of 520 µM [Na]₂-EDTA solution containing 0.2 U snake venom phosphodiesterase I, the sample was incubated for another 3 h at 37 °C. The total volume was 40 µl. The sample was then kept at –20 °C until the day of analysis. Samples were then filtered by using an AcroPrep Advance 96-filter plate 0.2 µm Supor (Pall Life Sciences) and then analysed by liquid chromatography electrospray ionization tandem mass spectrometry (LC-ESI-MS/MS), which are performed using an Agilent 1290 UHPLC system and an Agilent 6490 triple quadrupole mass spectrometer coupled with the stable isotope dilution technique. DNA samples were digested to give a nucleoside mixture and spiked with specific amounts of the corresponding isotopically labelled standards before LC-MS/MS analysis. The nucleosides were analysed in the positive ion selected reaction monitoring mode (SRM). In the positive ion mode, [M + H]⁺ species were measured.

Determination and comparison of nucleoside concentrations. The resulting cytosine, 5mC, 5hmC and 8-oxoG peak areas were normalized using the isotopically labelled standards, and expressed relative to the total cytosine content (that is, C + 5mC + 5hmC). Concentrations were depicted as averages of independent replicates grown on different days, and compared between hypoxia and normoxia (21% O₂), or between control and treated conditions, using a paired Student's *t*-test. No statistical methods were used to predetermine sample size.

RNA extraction, cDNA synthesis and qPCR. For RNA extraction, cell culture medium was removed, TRIzol (Life Technologies) added and processed according to manufacturer's guidelines. Reverse transcription and qPCR were performed using 2 × TaqMan Fast Universal PCR Master Mix (Life Technologies), TaqMan probes and primers (IDT, sequence in Supplementary Table 12). Thermal cycling

and fluorescence detection were done using a LightCycler 480 Real-Time PCR System (Roche). Taqman assay amplification efficiencies were verified using serial cDNA dilutions, and estimated to be >95%.

mRNA concentration analysis and statistics. Cycle threshold (C_t) values were determined for each sample and gene of interest in technical duplicates, and normalized according to the corresponding amplification efficiency. Per sample, *TET* expression was expressed relative to β -2-microglobulin (human) or hypoxanthine phosphoribosyltransferase 1 (*Hprt* mouse) levels by subtraction of their average C_t values. Concentrations were expressed as averages of at least 5 replicates extracted on different days. For Fig. 1a, copy number estimates for *TET1*, *TET2* and *TET3* were expressed for each cell line, relative to the summed copy number estimates of *TET1*, *TET2* and *TET3* under control conditions (21% O₂). Concentrations were compared between hypoxia and normoxia, or between control and treatment conditions using a Student's *t*-test. No statistical methods were used to predetermine sample size.

Hypoxia marker gene induction. To verify further induction of the hypoxia response program, hypoxia marker gene expression was verified. We analysed mRNA levels of genes encoding the E1B 19K/Bcl-2-binding protein Nip3 (*BNIP3*) and fructose-bisphosphate aldolase (*ALDOA*), 2 established hypoxia marker genes³³. Reverse transcriptase-quantitative PCR (RT-qPCR) was performed as described for the *TET* mRNA concentration assays, and differential expression was calculated using the $\Delta\Delta C_t$ method³⁴. We ruled out transcriptional upregulation as the cause of the increase in HIF1 α protein concentrations by assessing *HIF1A* mRNA expression in parallel. mRNA concentrations were expressed relative to normoxic controls (21% O₂). Differences in mRNA concentration were assessed using a Student's *t*-test on 5 or more independent replicates grown on different days.

Western blotting for Hif1 α , Tet1, Tet2 and Tet3. To assess Hif1 α protein stabilization, proteins were extracted from cultured cells as follows: cells were placed on ice, and washed twice with ice-cold PBS. Proteins were extracted with extraction buffer (50 mM Tris HCl, 150 mM NaCl, 1% Triton X-100, 0.5% sodium deoxycholate and 0.1% SDS) with 1 × protease inhibitor cocktail. Protein concentrations were determined using a bicinchoninic acid protein assay (BCA, Thermo Scientific) following the manufacturer's protocol. An estimated 60 µg protein was loaded per well on a NuPAGE Novex 3–8% Tris-Acetate Protein gel (Life Technologies), separated by electrophoresis and blotted on polyvinylidene fluoride membranes. Membranes were activated with methanol, washed and incubated with antibodies targeting β -actin (4967, Cell Signaling), Tet1 (09-872, Millipore) and Tet3 (61395, Active Motif), at 1:1,000 dilution, targeting Tet2 (124297, Abcam) at 1:250 dilution, and targeting Hif1 α (C-Term) (Cayman Chemical Item 10006421) at 1:3,000 dilution. Secondary antibodies and detection were according to routine laboratory practices. Western blotting was performed on 6 independent replicates grown on different days.

Analysis of HIF1 β target genes using ChIP-seq. To confirm that hypoxia-associated differential expression of *TET* genes is induced by the HIF pathway, we performed HIF1 β ChIP-seq. Because HIF1 β is the obligate binding partner of all three HIF α proteins stabilized and activated upon hypoxia³⁵, HIF1 β ChIP-seq reveals all direct HIF-target genes.

Chromatin immunoprecipitation. Approximately 25 × 10⁶–30 × 10⁶ cells were incubated in hypoxic conditions for 16 h. Cultured cells were subsequently immediately fixed by adding 1% formaldehyde (16% formaldehyde (w/v), Methanol-free, Thermo Scientific) directly to the medium and incubating for 8 min. Fixed cells were incubated with 150 µM of glycine for 5 min to revert cross-links, washed twice with ice-cold PBS 0.5% Triton X-100, scraped and collected by centrifugation (1,000g for 5 min at 4 °C). The pellet was re-suspended in 1,400 µl of RIPA buffer (50 mM Tris-HCl pH 8, 150 mM NaCl, 2 mM EDTA pH 8, 1% Triton X-100, 0.5% sodium deoxycholate, 1% SDS, 1% protease inhibitors) and transferred to a new Eppendorf tube. The lysate was homogenized by passing through an insulin syringe, and incubated on ice for 10 min. The chromatin was sonicated for 3 min by using a Branson 250 Digital Sonifier with 0.7 s 'On' and 1.3 s 'Off' pulses at 40% power amplitude, yielding a size of 100 to 500 bp. The sample was kept ice cold at all times during the sonication. The samples were centrifuged (10 min at 16,000g at 4 °C) and the supernatant was transferred in a new Eppendorf tube. The protein concentration was assessed using a BCA assay. Fifty microlitres of shared chromatin was used as 'input' and 1.4 µg of primary ARNT/HIF-1 β monoclonal antibody (NB100-124, Novus) per 1 mg of protein was added to the remainder of the chromatin, and incubated overnight at 4 °C in a rotator. Pierce Protein A/G Magnetic Beads (Life Technologies) were added to the samples in a volume four times the volume of the primary antibody and incubated at 4 °C for at least 5 h. A/G Magnetic Beads were collected and the samples were washed five times with the washing buffer (50 mM Tris-HCl, 200 mM LiCl, 2 mM EDTA, pH 8, 1% Triton, 0.5% sodium deoxycholate, 0.1% SDS, 1% protease inhibitors), and twice with a 10 mM Tris, 1 mM EDTA (TE) buffer. The A/G magnetic beads were re-suspended

in 50 µl of TE buffer, and 1.5 µl of RNase A (200 units, NEB) were added to the A/G beads samples and to the input, incubated for 10 min at 37 °C. After addition of 1.5 µl of proteinase K (200 U) and overnight incubation at 65 °C, the DNA was purified using 1.8 × volume of Agencourt AMPure XP (Beckman Coulter) according to the manufacturer's instructions, and then eluted in 15 µl of TE buffer. The input DNA was quantified on NanoDrop.

ChIP-seq, mapping and analysis. In total, 5 µg of input and all of the immunoprecipitated DNA was converted into sequencing libraries using the NEBNext DNA library prep master mix set. A single end of these libraries was sequenced for 50 bases on a HiSeq 2000, mapped using Bowtie and extended for the average insert size (250 bases). ChIP peaks were called by model-based analysis for ChIP-Seq³⁶, with standard settings and using a sequenced input sample as baseline.

Patient-derived xenografted tumours. To assess whether tumour-associated hypoxia reduces 5hmC levels *in vivo*, redundant material from two endometrial tumours and a breast tumour, removed during surgery, was grafted in the interscapular region of nude mice. Informed consent was obtained from the patient, following the ethical approval of the local ethical committee. All animal experiments were approved by the local ethical committee (P098/2014). Each tumour was allowed to grow to 1 cm³, after which it was collected. 10% of this tumour was re-implanted in a nude mouse, and the tumour was propagated for three generations until it was used for this experiment. To mark hypoxic areas, mice were injected with pimonidazole (60 mg kg⁻¹, Hypoxyprobe) i.p. 1 h before killing.

Immunofluorescence staining and analysis. Tumours were collected, fixed in formaldehyde and embedded in paraffin using standard procedures. Paraffin was removed and slides were rehydrated in two xylene baths (5 min), followed by five 3-min ethanol baths at decreasing concentrations (100%, 96%, 70%, 50% and water) and a 3-min TBS (50 mM Tris, 150 mM NaCl, pH 7.6) bath.

The following antibodies were used for immunofluorescence staining: primary antibodies were FITC-conjugated mouse anti-pimonidazole (HP2-100, Hydroxyprobe), rabbit anti-5hmC (39791, Active Motif), rat anti-polyoma middle T (AB15085, Abcam), rat anti-CD31 (557355, BD Biosciences), rat anti-CD45 (553076, BD Biosciences), rabbit anti-Ki67 (AB15580, Abcam) and mouse anti-pan cytokeratin (C2562, Sigma). Secondary antibodies were Alexa Fluor 405-conjugated goat anti-rabbit (A31556, Thermo Fisher), Alexa Fluor 647 conjugated goat anti-rat (A-21247, Life Technologies), peroxidase-conjugated goat anti-FITC (PA1-26804, Pierce), biotinylated goat anti-rat (A10517, Thermo Fisher) and biotinylated goat anti-rabbit (E043201, Dako). Signal amplification was performed using the TSA Fluorescein System (NEL701A001KT, Perkin Elmer) or the TSA Cyanine 5 System (NEL705A001KT, Perkin Elmer).

Different protocols were implemented depending on the epitopes of interest. Staining for the following epitopes was combined: CD45, 5hmC, pimonidazole and DNA; PyMT, 5hmC, pimonidazole and DNA; Ki67, pimonidazole and DNA; CD31 and pimonidazole; and pan-cytokeratin, 5hmC, pimonidazole and DNA.

Antigen retrieval for CD31, CD45 and pan-cytokeratin was done by a 7-min trypsin digestion, for pimonidazole and Ki67 using AgR at 100 °C for 20 min, followed by cooling for 20 min. Slides were washed in TBS for 5 min, permeabilized in 0.5% Triton X-100 in PBS for 20 min. For 5hmC antigen retrieval, slides were denatured in 2 M HCl for 10 min; HCl was neutralized for 2 min in borax, 1% in PBS pH 8.5, and washed twice for 5 min in PBS.

For all slides, endogenous peroxidase activity was quenched using H₂O₂ (0.3% in methanol), followed by three 5-min washes in TBS. Slides were blocked using pre-immune goat serum (X0907, Dako; 20% in TNB; TSA Biotin System kit, Perkin Elmer). Binding of primary antibodies (anti-5hmC, anti-CD45, anti-CD31 and anti-pan cytokeratin or FITC-conjugated anti-pimonidazole; all 1:100 in TNB) was allowed to proceed overnight. Slides were washed 3 times in TNT (0.5% Triton-X100 in TBS) for 5 min, after which the following secondary antibodies (all 1:100 in TNB with 10% pre-immune sheep serum) were allowed to bind for 45 min: sheep anti-FITC-PO (for pimonidazole), goat anti-rabbit-Alexa Fluor 405 (for 5hmC), goat anti-rat-Alexa Fluor 647 (for CD45), and biotinylated goat anti-mouse (for pan-cytokeratin). Slides were washed three times for 5 min in TNT, after which signal amplification was performed for 8 min using Fluorescein Tyramide (1:50 in amplification diluent).

Slides stained for pimonidazole that required co-staining for Ki67 or PyMT, or slides stained for pan-cytokeratin that required co-staining for pimonidazole were subjected to a second indirect staining for the latter epitopes. After 5 min of TNT and 5 min of TBS, slides were quenched again for peroxidase activity using H₂O₂ and blocked using pre-immune goat serum, prior to a second overnight round of primary antibody binding (anti-Ki67, FITC-anti-pimonidazole or anti-PyMT, all 1/100). The next day, three 5-min washes with TNT were followed by a 1-h incubation with a biotinylated goat anti-rabbit antibody (for Ki67) or goat anti-rat (for PyMT), another three 5-min washes with TNT, a 30-min incubation with peroxidase conjugated to streptavidin (for Ki67 and PyMT) or to anti-FITC (for pimonidazole), another three 5-min washes with TNT and signal amplification

for 8 min using, for pimonidazole, Fluorescein Tyramide and for others Cyanine 5 Tyramide (1:50 in amplification diluent). Slides were then stained with propidium iodide with RNase (550825; BD biosciences) for 15 min, washed for 5 min in PBS and mounted with Prolong Gold (Life Technologies).

Slides were imaged on a Nikon A1R Eclipse Ti confocal microscope. Three to five sections per slide were imaged, and processed using ImageJ. Nuclei were identified using the propidium iodide signal and nuclear signal intensities for Fluorescein and Cy3 (pimonidazole and 5hmC) measured. Analyses were exclusively performed on slide regions showing a regular density and shape of nuclei, in order to avoid inclusion of acellular or necrotic areas. The pimonidazole signal will also not stain necrotic/acellular areas³⁷, and was used to stratify viable cell nuclei into normoxic (pimonidazole negative) and hypoxic (pimonidazole positive) regions. The 5hmC signals in each population were compared using ANOVA. PyMT-negative and CD45-positive cells were counted directly. The fraction of pimonidazole and CD31-positive areas was directly quantified using ImageJ across ten images per slide.

Metabolite and protein extraction. For metabolite extractions, 12-well cell culture dishes were placed on ice and washed twice with ice-cold 0.9% NaCl, after which 500 µl of ice-cold 80% methanol was added to each well. Cells were scraped and 500 µl was transferred to a vial on ice. Wells were washed with 500 µl 80% methanol, which was combined with the initial cell extracts. The insoluble fraction was pelleted at 4 °C by a 10-min 21,000g centrifugation. The pellet (containing the proteins) was dried, dissolved in 0.2 N NaOH at 96 °C for 10 min and quantified using a bicinchoninic acid protein assay (BCA, Pierce), whereas the supernatant fraction was processed for metabolite profiling.

Derivation and measurement of metabolites. The supernatant fraction containing the metabolites was transferred to a new vial and dried in a Speedvac. The dried supernatant fraction was dissolved in 45 µl of 2% methoxyamine hydrochloride in pyridine and held for 90 min at 37 °C in a horizontal shaker, followed by derivatization through the addition of 60 µl of N-(tert-butyldimethylsilyl)-n-methyl-trifluoroacetamide with 1% tert-butyldimethylchlorosilane and a 60-min incubation at 60 °C. Samples were subsequently centrifuged for 5 min at 21,000g and 85 µl was transferred to a new vial and analysed using a gas-chromatography based mass spectrometer (triple quadrupole, Agilent) operated in Multiple Reaction Monitoring (MRM) mode.

Analysis of metabolite concentrations. For each sample, metabolite measurements were normalized per sample to the corresponding protein concentration estimates and expressed relative to control-treated samples. Four technical replicates were run for each sample, and the experiment was repeated 4 times using independent samples (*n* = 16). Differences in metabolite concentration were assessed using a two-tailed paired Student's *t*-test or using analysis of variance with post-hoc Tukey HSD when repeated measures were compared.

ROS measurement using 2',7'-dichlorodihydrofluorescein diacetate. MCF7 cells were cultured in 24-well plates and exposed to 21% (control) or 0.5% O₂ (hypoxia) for 24 h. DMEM used for staining was pre-equilibrated to the required O₂ tension, and all steps performed at 21% (control) or 0.5% O₂ (hypoxia) using a glove box. The cells were washed twice with 500 µl DMEM, and incubated for 30 min in 2',7'-dichlorodihydrofluorescein diacetate (DCF-DA; 10 µM) in 500 µl DMEM, keeping 2 wells unstained by DMEM without DCF-DA. Cells were treated with the indicated concentrations of H₂O₂ in DMEM for 30 min at 37 °C, and fixed by adding 33.3 µl of 16% methanol-free paraformaldehyde (Thermo Fisher) for 8 min at room temperature. The fixative was quenched using glycine (150 µM), cells were washed twice in ice-cold PBS, scraped to detach them and transfer them to pre-cooled FACS tubes over cell strainers. Cells were kept on ice until they were analysed by flow cytometry using a FACSVerse (BD Biosciences).

Nuclear ROS measurement using nuclear peroxy emerald 1. MCF7 cells were seeded on 12-well glass-bottom plates and after 24 h exposed to 21% (control) or 0.5% O₂ (hypoxia) for 24 h. PBS used for subsequent staining was pre-equilibrated to the required O₂ tension, and all washing, treatment and staining steps were performed at the appropriate O₂ tension (21% or 0.5%) using a glove box. Cells were loaded with nuclear peroxy emerald 1 (NucPE1; 5 µM)^{38,39} and Hoechst 33342 (10 µg ml⁻¹) in PBS for 15 min at 37 °C. After washing three times in PBS, control cells were incubated with H₂O₂ (0.5 mM in PBS) as a positive control, or with water (control and hypoxia cells) in PBS at 37 °C for 20 min. Cells were washed three times in PBS, placed on ice and immediately imaged by confocal microscopy. The nuclear NucPE1 signal was measured, and averaged across >100 nuclei per replicate using ImageJ. This experiment was repeated 5 times on different days, and signals compared using a *t*-test.

Cell growth measurement using Sulforhodamine B. 5,000 cells/well were seeded in three 96-well plates. After 48 h, one plate was fixed using trichloroacetic acid (3.3% w/v) for 1 h at 4 °C, one plate incubated for 24 h at 37 °C under hypoxic and one under control conditions (0.5% and 21% O₂, respectively). The latter 2 plates were subsequently also fixed using trichloroacetic acid (3.3% wt/vol) for 1 h at

4°C, and all 3 plates were next analysed using the *In vitro* Toxicology Assay Kit, Sulforhodamine B-based (Sigma) as per the manufacturer's instructions. Growth inhibition was calculated as described⁴⁰.

siRNA transfection. siRNA ON-TARGETplus SMART pools (Thermo) were diluted in OptiMEM I reduced serum medium using Lipofectamine RNAiMAX (Life technologies) to reverse-transfect MCF7 cells in 10-cm dishes (for DNA) or 6-well plates (for RNA). Cells were transfected 72 h before RNA and DNA extraction as described.

Hydroxylation assay using nuclear extracts. MCF7 cells were cultured for 24 h under control or hypoxic conditions (21% or 0.5% O₂, respectively), chilled on ice and processed for extraction of nuclear proteins using the NE-PER Nuclear and Cytoplasmic Extraction Kit (Thermo Scientific). The activity of control and hypoxic extracts was assessed in parallel using the Colorimetric Epigenase 5mC-Hydroxylase TET Activity/Inhibition Assay Kit (Epigentek) according to manufacturer's instructions. Reactions were allowed to proceed for one hour, after which washing and detection of 5hmC were done according to manufacturer's instructions. Differences between hypoxia and control were analysed using ANOVA, for 5 independent experiments.

DNA hydroxymethylation assay using purified Tet enzyme. The genomic DNA used in this assay was extracted from *Tet* triple-knockout ES cells (G. -L. Xu), and it therefore was devoid of 5hmC⁴¹. To enable efficient denaturation, it was digested using *MseI* before the assay and purified using solid phase reversible immobilisation paramagnetic beads (Agencourt AMPure XP, Beckman Coulter). The assays were performed in Whitley H35 Hypoxystation (don Whitley Scientific) at 37°C, 5% CO₂, N₂, with the following oxygen tensions: 0.1%, 0.3%, 0.5%, 1%, 2.5%, 5%, 10% and 21%. Hypoxystations were calibrated less than 1 month before all experiments. Optimized assay components were as follows: 1.0 µg µl⁻¹ bovine serum albumin (New England Biolabs), 50 mM Tris (pH 7.8), 100 µM dithiothreitol (Life Technologies), 2 ng µl⁻¹ digested gDNA, 250 µM α-ketoglutarate, 830 µM ascorbate, 200 µM FeSO₄ and 45 ng µl⁻¹ Tet1 enzyme (WiseGene). The major assay components (H₂O, BSA and Tris) used for all samples were allowed to pre-equilibrate at 0.1% O₂ for 1 h. These and the remaining assay buffer components (<100 µl) were then pre-equilibrated at the desired oxygen tension for 15 min, and mixed before addition of Tet1 enzyme in a total reaction volume of 25 µl. Reactions were allowed to proceed for 3 min, longer incubations showed a decrease in activity. Reactions were stopped with 80 mM EDTA and stored at -80°C. To measure the resulting 5hmC content of the DNA, reactions were diluted to 100 µl, denatured for 10 min at 98°C and analysed in duplicate using the Global 5-hmC Quantification Kit (Active Motif) following manufacturer's instructions. Michaelis-Menten and Lineweaver-Burk plots and the resulting *K_M* values were estimated using R.

Hypoxia-induced changes in genomic distribution of 5(h)mC in MCF7 cells: DIP-seq. To assess where in the genome the levels of 5mC and 5hmC were altered, we performed DNA immunoprecipitations coupled to high-throughput sequencing (DIP-seq). MCF7 cells were selected for these experiments as they were a cancer cell line with high levels of 5hmC and expression of *TET* genes under control conditions, and a cell growth that is unaffected by hypoxia. This enabled us to study the effects of hypoxia on TET activity in a cell line that shows high endogenous activity, but that is isolated from hypoxia-induced changes in cell proliferation. MCF7 cell culture and DNA extractions were as described for LC-MS analyses. Library preparations and DNA immunoprecipitations were performed as described⁴², using established antibodies targeting 5mC (clone 33D3, Eurogentec,) and 5hmC (Active Motif catalogue number 39791). For 5hmC-DIP-seq, paired barcoded libraries prepared from DNA of hypoxic and control samples were mixed before capture, to enable a direct comparison of 5hmC-DIP-seq signal to the input. A single end of these libraries was sequenced for 50 bases on a HiSeq 2000, mapped using Bowtie and extended for the average insert size (150 bases). Mapping statistics are summarized in Supplementary Information Table 11.

For analysis of sequencing data, MACS peak calling, read depth quantification and annotation with genomic features as annotated in Ensembl build 77 was performed using SeqMonk. Differential (hydroxy-)methylation was quantified by EdgeR⁴³, using either 3 or 5 independent pairs of control and hypoxic samples (for 5hmC-DIP-seq and 5mC-DIP-seq, respectively). These cells were cultured and exposed to hypoxia (0.5% O₂) or control conditions (21% O₂) on different days. Results were reported for 5hmC peak areas that exhibited a change significant at a *P* < 0.05 and 5% FDR.

Target enrichment BS-seq using SeqCapEpi. To confirm enrichment of 5mC at gene promoters using an independent method, DNA libraries were prepared using methylated adapters and the NEBNext DNA library prep master mix set following manufacturer recommendations. Libraries were bisulfite-converted using the Imprint DNA modification kit (Sigma) as recommended, and PCR amplified for 12 cycles using barcoded primers (NEB) and the KAPA HiFi HS Uracil+ ready mix (Sopachem) according to manufacturer's instructions. Fragments were selected from these libraries using the SeqCapEpi CpGiant Enrichment Kit (Roche)

following the manufacturer's instructions, sequenced from both ends for 100 bases on a HiSeq 2000.

For analysing these sequences, sequencing reads were trimmed for adapters using TrimGalore and mapped on a bisulfite-converted human genome (GRCh37) using BisMark. The number of methylated and un-methylated cytosines in captured regions was quantified using Seqmonk for each experiment. Differential methylation of regions of interest was assessed by Fisher's exact test and for 5 independent replicates grown on different days. *t*-scores were averaged following Fisher's method. Mapping statistics are summarized in Supplementary Table 11.

RNA-seq. To assess the effect of the increased 5mC occupancy at gene promoters on their expression, RNA-seq was performed. Briefly, total RNA was extracted using TRIzol (Invitrogen), and remaining DNA contaminants in 17–20 µg of RNA was removed using Turbo DNase (Ambion) according to the manufacturer's instruction. RNA was repurified using RNeasy Mini Kit (Qiagen). Ribosomal RNA present was depleted from 5 µg of total RNA using the RiboMinus Eukaryote System (Life technologies). cDNA synthesis was performed using SuperScript III Reverse Transcriptase kit (Invitrogen). 3 µg of Random Primers (Invitrogen), 8 µl of 5× First-Strand Buffer and 10 µl of RNA mix was incubated at 94°C for 3 min and then at 4°C for 1 min. 2 µl of 10 mM dNTP Mix (Invitrogen), 4 µl of 0.1 M DTT, 2 µl of SUPERase In RNase Inhibitor 20U µl⁻¹ (Ambion), 2 µl of SuperScript III RT (200 U µl⁻¹) and 8 µl of Actinomycin D (1 µg µl⁻¹) were then added and the mix was incubated for 5 min at 25°C, 60 min at 50°C and 15 min at 70°C to heat-inactivate the reaction. The cDNA was purified by using 80 µl (2× volume) of Agencourt AMPure XP and eluted in 50 µl of the following mix: 5 µl of 10× NEBuffer 2, 1.5 µl of 10 mM dNTP mix (10 mM dATP, dCTP, dGTP, dUTP, Sigma), 0.1 µl of RNaseH (10 U µl⁻¹, Ambion), 2.5 µl of DNA Polymerase I Klenov (10 U µl⁻¹, NEB) and the remaining volume of water. The eluted cDNA was incubated for 30 min at 16°C, purified by Agencourt AMPure XP and eluted in 30 µl of dA-Tailing mix (2 µl of Klenow Fragment, 3 µl of 10× NEBNext dA-Tailing Reaction Buffer and 25 µl of water). After 30 min incubation at 37°C, the DNA was purified by Agencourt AMPure XP, eluted in TE buffer and quantified on NanoDrop. Subsequent library preparation was performed using the DNA library prep master mix set and sequencing was performed as described for ChIP-seq. Expression levels (reads per million) of genes displaying significant increases in methylation at their gene promoter, as determined using SeqCapEpi, was compared between control and hypoxic samples using a *t*-test. Mapping statistics are summarized in Supplementary Table 11.

TCGA samples and data analysis. From the TCGA pan-cancer analysis, we selected all solid tumour types for which >100 tumours were available with both gene expression data (RNA-seq) and DNA methylation data (Illumina Infinium HumanMethylation450 BeadChip). These were 408 bladder carcinomas, 691 breast carcinomas, 243 colorectal adenocarcinomas, 520 head and neck squamous cell carcinomas, 290 kidney renal cell carcinomas, 430 lung adenocarcinomas, 371 lung squamous cell carcinomas, and 188 uterine carcinomas, representing in total 3,141 unique patients. Corresponding RNA-seq read counts as well as DNA methylation data from Infinium HumanMethylation450 BeadChip arrays were downloaded from the TCGA server. Breast tumour subtypes were annotated for 208 tumours and, for the remaining tumours, imputed by unsupervised hierarchical clustering of genes in the PAM50 gene expression signature⁴⁴. Other clinical and histological variables were available for >95% of tumours, and missing values were encoded as not available. Gene mutation data was available for 129 bladder carcinomas, 646 breast carcinomas, 200 colorectal adenocarcinomas, 306 head and neck squamous cell carcinomas, 241 kidney renal cell carcinomas, 182 lung adenocarcinomas, 74 lung squamous cell carcinomas, and 3 uterine carcinomas.

Stratification of tumours for hypoxia and proliferation. To identify which of these tumour samples were hypoxic or normoxic, we performed unsupervised hierarchical clustering based a modification (Ward.D of the clusth function in R's stats package) of the Ward error sum of squares hierarchical clustering method⁴⁵, on normalized log₁₀-transformed RNA-seq read counts for 14 genes that make up the hypoxia metagene signature (*ALDOA*, *MIF*, *TUBB6*, *P4HA1*, *SLC2A1*, *PGAM1*, *ENO1*, *LDHA*, *CDKN3*, *TPI1*, *NDRG1*, *VEGFA*, *ACOT7* and *ADM*)²⁵. In each case the top 3 sub-clusters identified were annotated as normoxic, intermediate and hypoxic. To identify which of these tumour samples were high- or low-proliferative, we performed unsupervised hierarchical clustering based a modification (Ward.D of the clusth function in R's stats package) of the Ward error sum of squares hierarchical clustering method⁴⁵, and this for all genes annotated to an established tumour proliferation signature (*MKI67*, *NDC80*, *NUF2*, *PTTG1*, *RRM2*, *BIRC5*, *CCNB1*, *CEP55*, *UBE2C*, *CDC20* and *TYMS*)⁴⁶. Tumours in the top 2 sub-clusters identified were labelled as high- or low-proliferative.

Analysis of the top 1000 CpGs most hypermethylated versus normal tissue. To identify tumour-associated hypermethylation events, we compared 450k methylation data from tumours and normal tissues. All available DNA methylation data from normal tissue (matched or unmatched to tumour samples, on average

59 per tumour type, representing 472 in total, range = 21–160) were downloaded. For each of the 8 tumour types investigated, we selected the top 1,000 CpGs that showed the highest average tumour-associated increases in DNA methylation. Per tumour type, unsupervised hierarchical clustering based on a modification of the Ward error sum of squares hierarchical clustering method (Ward.D of the *clusth* function in R's stats package)⁴⁵ annotated the first 3 clusters identified as having low, intermediate and high hypermethylation. Cluster co-membership for methylation and hypoxia metagene expression were analysed using the Cochran–Armitage test for trend. Analyses using the top 100, 500, 5,000 or 10,000 CpGs yielded near identical results (not shown).

Analysis of hypermethylation events. We next applied a method to identify those CpGs that exhibit exceptional increases in hypermethylation but that are hypermethylated only in a subset of all tumours. Such rare events are typically found in cancer, where hypermethylation inactivates a gene in only a subset of tumours. Hypermethylation of individual CpGs at gene promoters (that is, on average 3.7 CpGs per promoter are represented on the 450K array) in individual tumours was assessed as follows: To achieve a normal distribution, all β -values were transformed to M-values⁴⁷ using $M = \log_2(\beta/(1 - \beta))$. For each tumour type, the mean μ and standard deviation σ of the M value across all control (normoxic) tumours was next calculated, irrespective of mutational status, for each CpG, and used to assign Z-values to each CpG in each tumour using $Z = (M - \mu)/\sigma$. These Z-values describe the deviation in normal methylation variation for that probe. To identify CpGs that display an extreme deviation, we selected those for which the Z-value exceeded 5.6 (that is, $\mu + (5.6 \times \sigma)$), corresponding to a Bonferroni-adjusted *P* value of 0.01; they were considered as hypermethylation events in that particular tumour. This analysis was preferred over Wilcoxon-based models that assess differences in the average methylation level between subgroups, as the latter do not enable the identification or quantification of the rarer hypermethylation events in individual promoters or CpGs.

To identify genes with frequently hypermethylated CpGs in their promoter, the number of hypermethylation events in that promoter was counted in all tumours, and contrasted to the expected number of hypermethylation events in that promoter (that is, the general hypermethylation frequency multiplied by the number of CpGs assessed in that promoter multiplied by the number of tumours) using Fisher's exact test. Genes with an associated Bonferroni-adjusted *P* value below 0.01 were retained and considered as frequently hypermethylated in that tumour type.

To assess what fraction of these hypermethylation events are hypoxia-related, we assumed that the fraction of events detected under normoxia was hypoxia-unrelated, and that all excess events detected in intermediate and hypoxic tumours were hypoxia-related. For example, in 691 breast carcinomas, 0.25% of CpGs were hypermethylated in 251 normoxic tumours, 0.81% in 350 intermediate and 1.40% in 90 hypoxic tumours. So, 0.56% and 1.15% of hypermethylation events in respectively intermediate and hypoxic tumours were hypoxia-related. Taking into account the number of tumours, 0.25% of hypermethylation events (that is, $(0.25\% \times 251 + 0.25\% \times 350 + 0.25\% \times 90)/691$) are not hypoxia-related, and 0.43% are hypoxia related (that is, $(0\% \times 251 + 0.56\% \times 350 + 1.15\% \times 90)/691$). So, 63% of all hypermethylation events combined (that is, $0.43/(0.43 + 0.25)$) are hypoxia related. To assess the contribution of hypoxia to hypermethylation relative to other covariates, partial R^2 values were calculated for the contribution of each covariate in a linear model, and compared to the total R^2 achieved by the model.

To identify genes with CpGs in their promoter that are more frequently hypermethylated in hypoxic than normoxic tumours, the number of hypermethylation events in that promoter was counted in all hypoxic tumours, and contrasted to the number found in normoxic tumours. Differences in frequencies were assessed using Fisher's exact test, and genes with a Bonferroni-adjusted *P* < 0.01 were retained and considered hypermethylated upon hypoxia. Enrichment of ontologies associated with these genes was assessed using Fisher's exact test as implemented in R's topGO package.

Analysis of the effect of hypermethylation events on the expression of associated genes. To enable a direct comparison between the expression of different genes, we transformed gene expression values (reads per million) to their respective z-scores. To assess the impact of hypermethylation events on the expression of associated genes, we compared the expression z-scores of all frequently hypermethylated genes that contain one or more hypermethylation events in their promoter (on average each promoter contains 3.7 CpGs; if one of these is hypermethylated the associated gene is considered hypermethylated in that particular tumour), to the expression of all frequently hypermethylated genes that do not contain a hypermethylation event. The effect of hypermethylation on gene expression was plotted for the 8 main tumour types stratified into normoxic, intermediately hypoxic and hypoxic tumours, and for glioblastomas was stratified into normoxic, intermediately hypoxic, hypoxic and *IDH*-mutant tumours (*n* = 4). The difference in expression z-scores between genes not carrying and carrying a hypermethylation event in their promoter was assessed using a *t*-test.

Analysis of the effect of frequent mutations on the occurrence of hypermethylation events. To assess the impact of somatic mutations on hypoxia-associated hypermethylation frequencies, we analysed the top 20 genes described to be most frequently mutated in the pan-cancer analysis²⁴, and supplemented this list with genes known to cause hypermethylation upon mutation (that is, *IDH1*, *IDH2*, *SDHA*, *FH*, *TET1*, *TET2* and *TET3*). Mutations in *IDH1* and *IDH2* were retained if they respectively affected amino acid R132, and amino acids R140 or R172. Mutations in other genes were scored using Polyphen, and only mutations classified as probably damaging were retained. 7 mutations were found in lung tumours, 3 mutations in colorectal tumours, 8 mutations in breast tumours and 6 mutations (all *IDH1*^{R132}) in glioblastomas. None of these mutations were enriched in hypoxic subsets. In multivariate analyses of variance, in each of the tumour types analysed, mutations in these genes were significantly associated with increased hypermethylation frequencies. Hypoxia was independently and significantly associated with the hypermethylation frequency.

Correlation between hypermethylation and expression of TET or DNMT enzymes. Gene expression values (reads per million) of DNMT and TET enzymes were determined for each tumour using available RNA-seq data. The number of hypermethylation events at significantly hypermethylated genes in each tumour was determined as described above. Hypermethylation in each tumour was subsequently correlated to *TET* or *DNMT* gene expression in that tumour, correcting for hypoxia and proliferation status using ANOVA.

5mC and 5hmC profiling using 450k arrays for 24 lung tumours. Newly diagnosed and untreated non-small-cell lung cancer patients scheduled for curative-intent surgery were prospectively recruited. Included subjects had a smoking history of at least 15 pack-years. The study protocol was approved by the Ethics Committee of the University Hospital Gasthuisberg (Leuven, Belgium). All participants provided written informed consent. In the framework of a different project⁴⁸, RNA-seq was performed on 39 tumours from these patients. Gene expression for these samples was clustered for their hypoxia metagene signature²⁵. This yielded 2 clear clusters, containing 24 and 15 normoxic and hypoxic tumours, respectively. Twelve samples were randomly selected from each cluster, in order to perform 5hmC and 5mC profiling.

Illumina Infinium HumanMethylation450 BeadChips. For TAB–ChIP, DNA was glycosylated and oxidized as described⁴⁹, using the 5hmC TAB–Seq Kit (WiseGene). Subsequently, bisulfite conversion, DNA amplification and array hybridization were done following manufacturer's instructions.

Analysis of TAB–ChIP and BS–ChIP. Data processing was largely as described⁵⁰. In brief, intensity data files were read directly into R. Each sample was normalized using Subset–quantile within array normalization (SWAN) for Illumina Infinium HumanMethylation450 BeadChips⁴⁹. Batch effects between chips and experiments were corrected using the runComBat function from the ChAMP bioconductor package⁵¹. For obtaining 5mC-specific beta values, TAB–ChIP generated normalized beta values were subtracted from the standard 450K generated normalized beta values, exactly as described⁵⁰.

Mouse cancer models. All the experimental procedures were approved by the Institutional Animal Care and Research Advisory Committee of the KU Leuven.

Hypoxia induction using sFlk1-overexpression. For sFlk1-overexpression studies, male Tg(MMTV–PyMT) FVB mice were intercrossed with wild-type FVB female mice. Female pups of the Tg(MMTV–PyMT) genotype were retained, and tumours allowed to develop for 9 weeks. Subsequently, 2.5 µg of plasmid (sFlk1-overexpressing or empty vector; randomly assigned within litter mates) per gram of mouse body weight was introduced in the bloodstream using hydrodynamic tail vein injections⁵². sFlk1 overexpression was monitored at 4 days after injection and at the day of killing (18 days after the injection), by eye bleeds followed by an ELISA assay for sFlk1 (R&D Systems) in blood plasma. At 12 weeks of age, mice were killed and mammary tumours collected were blinded for treatment.

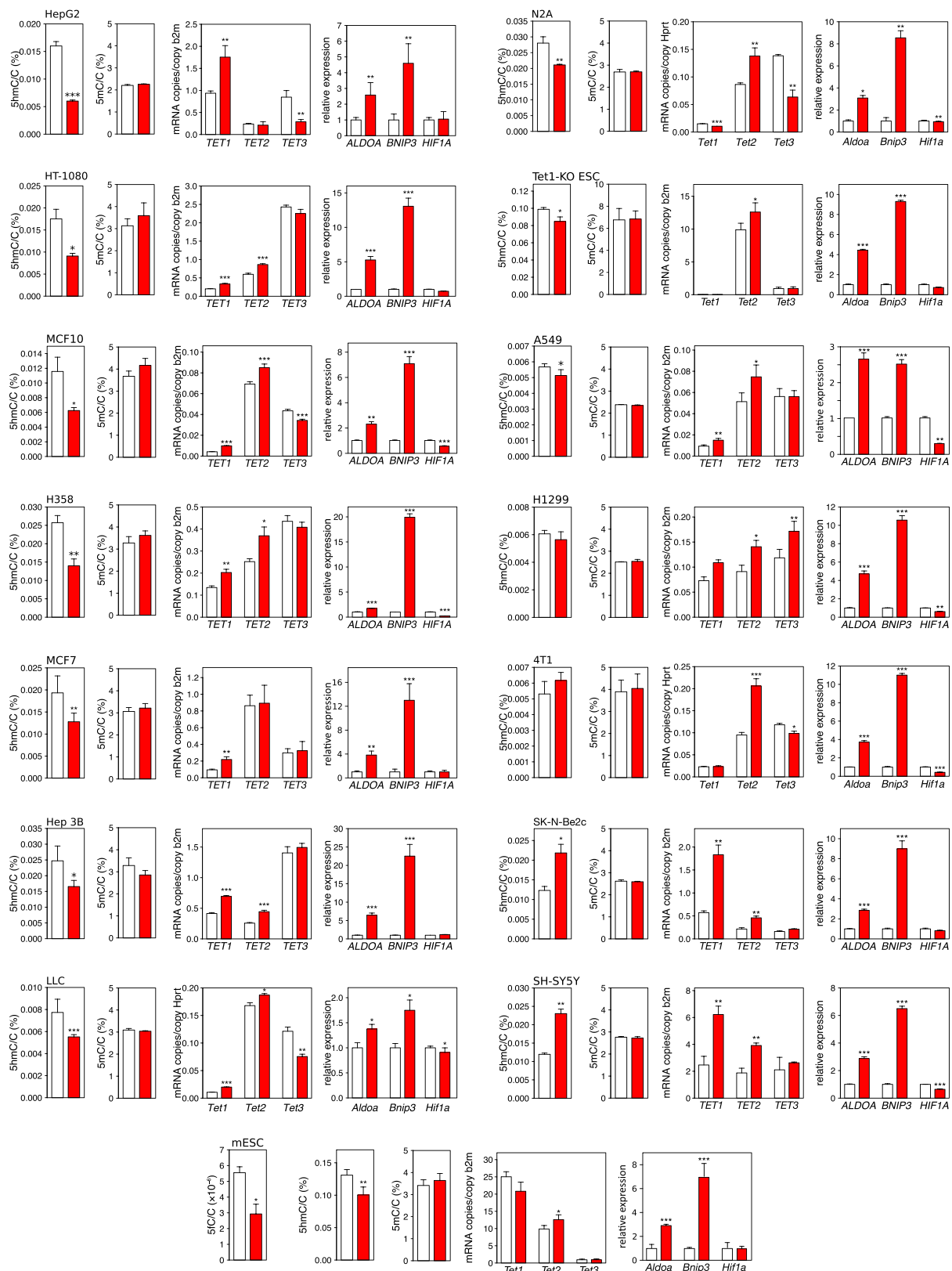
Hypoxia reduction using Phd2 haploinsufficiency. For the *Phd2*^{+/−} experiments, male Tg(MMTV–PyMT) FVB mice were intercrossed with female *Phd2*^{−/−} mice, yielding litters of which half have either a Tg(MMTV–PyMT) genotype or a Tg(MMTV–PyMT);*Phd2*^{−/−} genotype. For the *Phd2*^{WT/Δ} experiments, male Tg(MMTV–PyMT) FVB mice were intercrossed with female *Tie2*–Cre;*Phd2*^{WT/Δ} mice as described²⁷, yielding litters of which half have either a Tg(MMTV–PyMT);*Phd2*^{WT/WT} genotype or a *Tie2*–cre;Tg(MMTV–PyMT);*Phd2*^{−/−} genotype. At 16 weeks of age, female mice were killed and mammary tumours collected.

qPCR analysis of expression of Tets and marker genes. RNA was extracted from fresh-frozen tumours (stored at −8 °C) using TRIzol (Life Technologies), and converted to cDNA and quantified as described for the cell lines. TaqMan probes and primers (IDT or Life Technologies) are described under Supplementary Table 12. **TAB–seq of PyMT tumours.** TAB–seq libraries were prepared as described, using the 5hmC TAB–Seq Kit (WiseGene). DNA was bisulfite-converted using the EZ DNA Methylation–Lightning Kit (Zymo Research) and sequenced as described for SeqCapEpi experiments. Reads were mapped to the mouse genome

(build Mm9) and further data processing was as for SeqCapEpi experiments. DNA from 3 independent tumours was selected per condition. TET oxidation efficiency was required to exceed 99.5% as estimated using a fully CG-methylated plasmid spike-in, 5hmC protection by glycosylation was 65% as estimated using a fully hydroxymethylated plasmid spike-in, bisulfite conversion efficiencies were estimated to exceed 99.8% based on nonCG methylation (equal to percentage hypermethylated CpG). Mapping statistics are summarized in Supplementary Table 11. **Targeted deep BS-seq.** As no mouse capture kit was available for targeted BS-seq, a specific ampliconBS was developed for a set of 15 tumour suppressor gene promoters and 5 oncogene promoters. More specifically, DNA was bisulfite-converted using the Imprint DNA modification kit and amplified using the MegaMix Gold 2× mastermix and validated primer pairs. Per sample, PCR products were mixed to equimolar concentrations, converted into sequencing libraries using the NEBNext DNA library prep master mix set and sequenced to a depth of ~500×. Mapping and quantification were done as described for SeqCapEpi. The average and variance of methylation level M values in normal mammary glands were used as baseline, and amplicons displaying over 3 standard deviations more methylation (FDR-adjusted $P < 0.05$) than this baseline were called as hypermethylated. At least 9 different tumours, each from different animals, were profiled per genotype or treatment, and differences in hypermethylation frequencies between sets of tumours were assessed using Mann–Whitney's U-test.

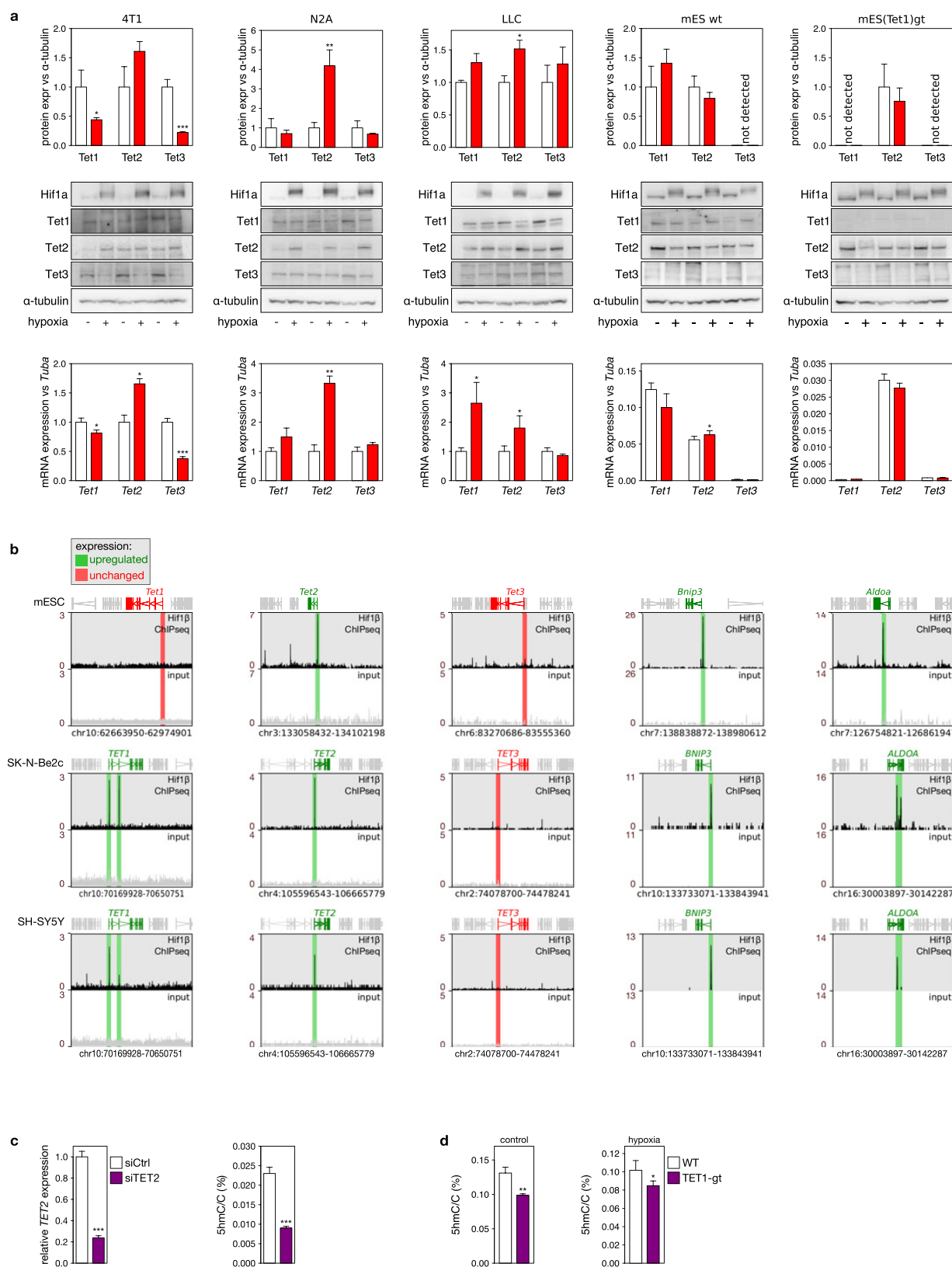
Statistics. Data entry and analysis were performed in a blinded fashion. Statistical significance was calculated by two-tailed unpaired *t*-test (Excel) or analysis of variance (R) when repeated measures were compared. Data were tested for normality using the D'Agostino–Pearson omnibus test (for $n > 8$) or the Kolmogorov–Smirnov test (for $n \leq 8$) and variation within each experimental group was assessed. Data are presented as mean \pm s.e.m. DNA methylation and RNA-seq gene expression data distributions were transformed to a normal distribution by conversion to M values and \log_2 transformation respectively. Sample sizes were chosen based on prior experience for *in vitro* and mouse experiments, or on sample and data availability for human tumour analyses. Other statistical methods (mostly related to specific sequencing experiments) are described together with the experimental details in other sections of the methods.

33. Sermeus, A. *et al.* Hypoxia induces protection against etoposide-induced apoptosis: molecular profiling of changes in gene expression and transcription factor activity. *Mol. Cancer* **7**, 27 (2008).
34. Schmittgen, T. D. & Livak, K. J. Analyzing real-time PCR data by the comparative C(T) method. *Nat. Protocols* **3**, 1101–1108 (2008).
35. Majmundar, A. J., Wong, W. J. & Simon, M. C. Hypoxia-inducible factors and the response to hypoxic stress. *Mol. Cell* **40**, 294–309 (2010).
36. Feng, J., Liu, T. & Zhang, Y. Using MACS to identify peaks from ChIP-Seq data. *Curr. Protoc. Bioinformatics* **Chapter 2**, Unit 2.14, (2011).
37. Durand, R. E. & Raleigh, J. A. Identification of nonproliferating but viable hypoxic tumor cells *in vivo*. *Cancer Res.* **58**, 3547–3550 (1998).
38. Lippert, A. R., Van de Bittner, G. C. & Chang, C. J. Boronate oxidation as a bioorthogonal reaction approach for studying the chemistry of hydrogen peroxide in living systems. *Acc. Chem. Res.* **44**, 793–804 (2011).
39. Dickinson, B. C., Tang, Y., Chang, Z. & Chang, C. J. A nuclear-localized fluorescent hydrogen peroxide probe for monitoring sirtuin-mediated oxidative stress responses *in vivo*. *Chem. Biol.* **18**, 943–948 (2011).
40. Vichai, V. & Kirtikara, K. Sulforhodamine B colorimetric assay for cytotoxicity screening. *Nat. Protocols* **1**, 1112–1116 (2006).
41. He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303–1307 (2011).
42. Taiwo, O. *et al.* Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat. Protocols* **7**, 617–636 (2012).
43. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
44. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).
45. Murtagh, F. & Legendre, P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *J. Classif.* **31**, 274–295 (2014).
46. Nielsen, T. O. *et al.* A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clin. Cancer Res.* **16**, 5222–5232 (2010).
47. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).
48. Wauters, E. *et al.* DNA methylation profiling of non-small cell lung cancer reveals a COPD-driven immune-related signature. *Thorax* **70**, 1113–1122 (2015).
49. Yu, M. *et al.* Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nat. Protocols* **7**, 2159–2170 (2012).
50. Nazor, K. L. *et al.* Application of a low cost array-based technique - TAB-Array - for quantifying and mapping both 5mC and 5hmC at single base resolution in human pluripotent stem cells. *Genomics* **104**, 358–367 (2014).
51. Morris, T. J. *et al.* ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics* **30**, 428–430 (2014).
52. Liu, F., Song, Y. & Liu, D. Hydrodynamics-based transfection in animals by systemic administration of plasmid DNA. *Gene Ther.* **6**, 1258–1266 (1999).
53. Pelicano, H., Carney, D. & Huang, P. ROS stress in cancer cells and therapeutic implications. *Drug Resist. Updat.* **7**, 97–110 (2004).
54. Booth, M. J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934–937 (2012).



Extended Data Figure 1 | Hypoxia-induced changes in 5hmC, 5mC and *TET* expression. Global 5hmC/C and 5mC/C content of DNA, *TET1*, *TET2* and *TET3* mRNA expression and hypoxia marker gene expression in 15 cell lines grown for 24 h under normoxic (21% O₂, white) or hypoxic (0.5% O₂, red) conditions. *TET* mRNA copy number is expressed relative to *B2M* for human cell lines (HepG2, HT-1080, MCF10A, H358, MCF7, Hep3B, A549, H1299, SK-N-Be2c and SHSY5Y), and to *Hprt* for mouse cell lines (LLC, N2A, 4T1, mES and Tet1-KO ES cells). Shown are cell lines derived from liver cancer (HepG2 and Hep3B), lung cancer (H358, A549, H1299 and LLC),

breast cancer (MCF7 and 4T1), fibrosarcoma (HT1080), neuroblastoma (SK-N-Be2c and SHSY5Y), normal breast epithelium (MCF10A) and the inner cell mass of blastocyst-stage mouse embryos (mES and Tet1-KO ES cells). *ALDOA* and *BNIP3* are expected to be increased, and *HIF1A* to be decreased upon hypoxia. The global 5fC content of ES cells is depicted, but was undetectable in cancer cell lines. Bars represent the mean \pm s.e.m. of five different replicate samples. DNA and RNA from these replicates were extracted from cells derived from the same stock vial but grown on different days. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ by paired *t*-tests.

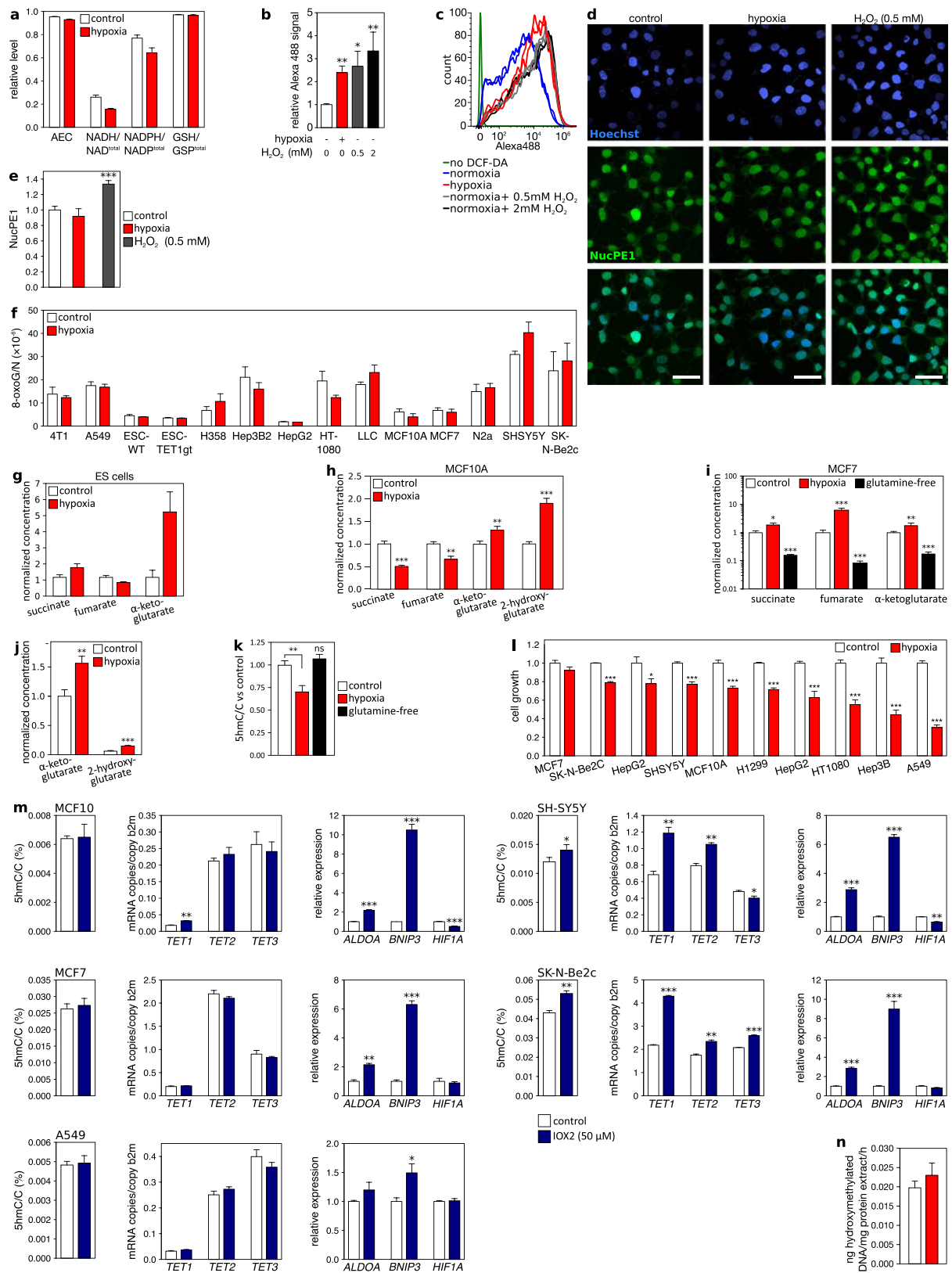


Extended Data Figure 2 | See next page for caption.

Extended Data Figure 2 | Impact of hypoxia on TET expression.

a, Changes in Tet1, Tet2 and Tet3 expression in mouse cell lines, at the protein level (top row, $n = 6$) and the mRNA level (bottom row, $n = 5$). Middle row: representative immunoblot images of Hif1a, Tet1, Tet2 and Tet3. α -Tubulin serves as loading control, and expression of the corresponding coding gene (*Tuba1a*) was used to normalize mRNA expression, enabling a direct comparison of relative protein and relative mRNA expression changes. For the same reason, mRNA expression was depicted relative to control conditions, in contrast to the absolute levels shown in Extended Data Fig. 1. Changes in Tet mRNA and protein expression correlate strongly (Pearson's $R = 0.855$, $P = 4 \times 10^{-4}$). For example, both 4T1 and N2A cells displayed increased Tet2 expression at the protein and mRNA level. Likewise, ES cells showed no pronounced changes at the protein or mRNA level. The overall expression of Tet enzymes was not altered in any of these cell lines. For gel source data,

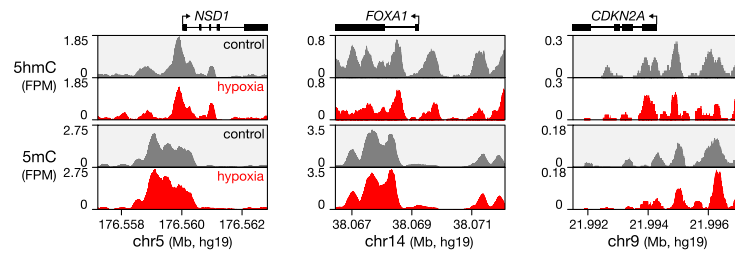
see Supplementary Fig. 1. **b**, Hif1 β ChIP-seq at the promoters of *TET1*, *TET2* and *TET3* and at hypoxia markers genes (*Bnip3* and *Aldoa*), with peaks or promoter regions highlighted using coloured boxes. Green and red boxes correspond to overexpression and no overexpression (specified in the figure panel) of the corresponding gene, respectively, as determined using TaqMan in Extended Data Fig. 1. Scale: reads per million reads and per base pair. **c**, Left, *TET2* expression in MCF7 cells transfected with control (white) or *TET2*-targeting (purple) siRNAs. Right, corresponding 5hmC levels as determined using LC-MS. **d**, 5hmC levels as determined by LC-MS, in wild-type (white) and *Tet1*-knockout (purple) ES cells grown under 21% (left) and 0.5% (right) O₂ tensions. Bars in **c** and **d** represent the mean \pm s.e.m. of five replicate samples from cells derived from the same stock vial but grown on different days. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ by paired t -tests (**a**, **c**, **d**).



Extended Data Figure 3 | See next page for caption.

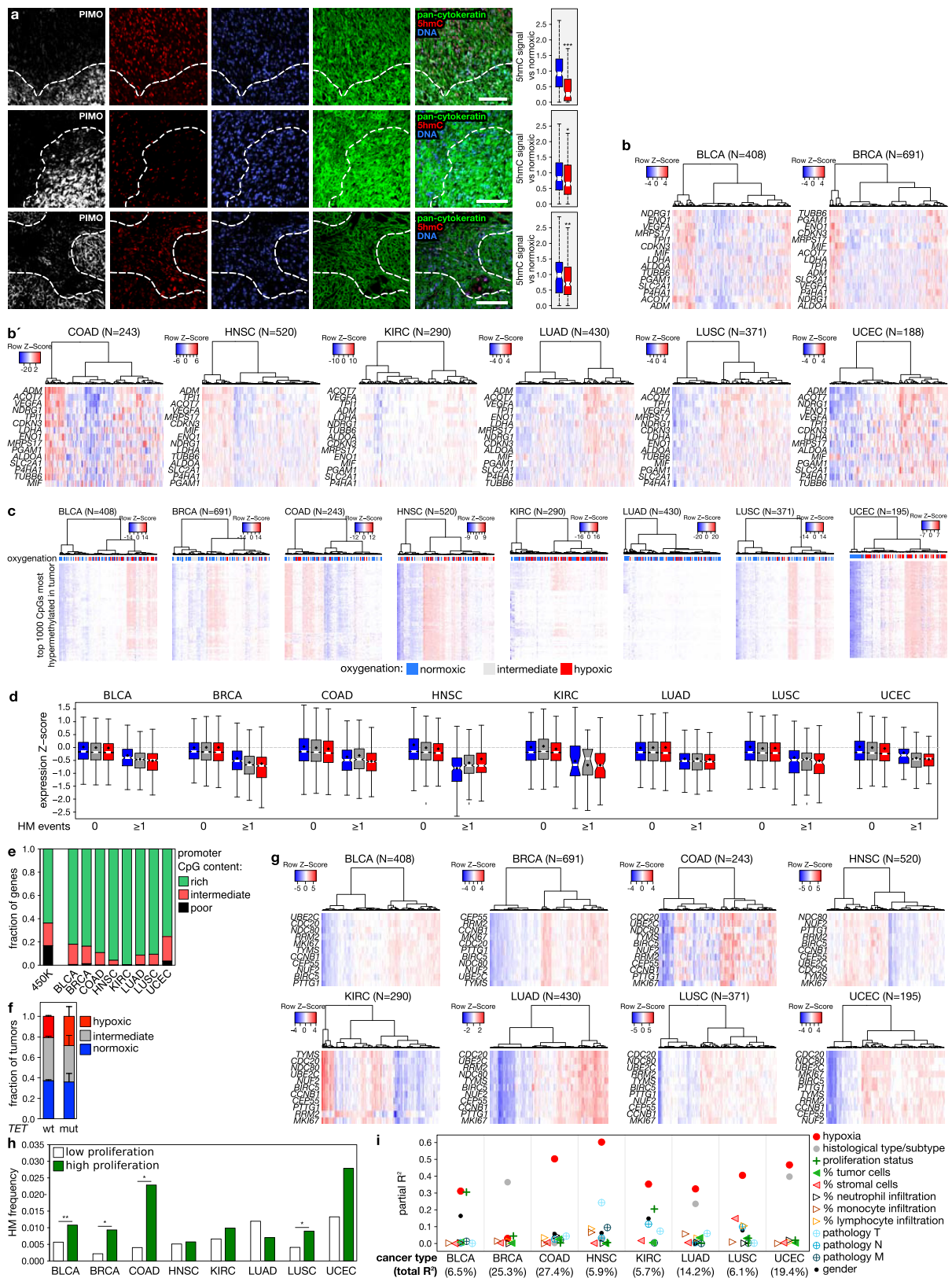
Extended Data Figure 3 | Effects secondary to hypoxia. a–e, ROS production and redox state of MCF7 cells cultured for 24 h under control (21% O₂, white) or hypoxic (0.5% O₂, red) conditions. Shown are capillary gas chromatography mass spectrometry (GC–MS) quantifications of changes in the cellular energy state as represented by the adenylate energy charge (AEC) (calculated as $[\text{ATP} + 0.5 \times \text{ADP}]/[\text{ATP} + \text{ADP} + \text{AMP}]$) (**a**); the reducing equivalents of the cell as represented by the relative NADH and NADPH levels (calculated as $\text{NADH}/[\text{NAD}^+ + \text{NADH}]$ and $\text{NADPH}/[\text{NADP}^+ + \text{NADPH}]$); and the reductive capacity of the cell as represented by the levels of glutathione (calculated as $\text{GSH}/[\text{GSH} + \text{GSSG} \times 2]$). **b, c,** Quantification (**b**) and representative FACS intensity traces (**c**) of total ROS levels in MCF7 cells exposed to hypoxia or H₂O₂, as assessed using 2',7'-dichlorodihydrofluorescein diacetate (DCF-DA). **d,** Nuclear ROS in MCF7 cells as assessed using the nuclear peroxyl emerald 1 probe (NucPE1)³⁹. MCF7 cells were exposed to 21% (control) or 0.5% (hypoxia) O₂ for 24 h, after which live cells were loaded with NucPE1 (5 μM) and Hoechst 33342 (10 μg ml⁻¹) in O₂ pre-equilibrated PBS for 15 min. After washing, control cells were incubated with H₂O₂ (0.5 mM in PBS) as a positive control, or with water (control and hypoxia cells) in PBS for 20 min. Cells were washed again and immediately imaged by confocal microscopy. Representative images are shown. Scale bar, 50 μm. **e,** The nuclear NucPE1 signal, averaged across >100 nuclei and expressed relative

to control conditions. **f,** LC–MS quantification of 8-oxoG concentrations in DNA of cells lines cultured for 24 h under control (21% O₂, white) and hypoxic (0.5% O₂, red) conditions. 8-oxoG serves as a marker of nuclear ROS⁵³. **g–i,** GC–MS quantification of changes in the indicated metabolite levels in mouse ES cells (**g**), MCF10A cells (**h**) and MCF7 cells (**i**) grown for 24 h under control (21% O₂, white), hypoxic (0.5% O₂, red) or glutamine-free (21% O₂, black) conditions. **j,** Quantities of α-ketoglutarate and 2-hydroxyglutarate in MCF7 cells, expressed relative to α-ketoglutarate levels in MCF7 cells grown under control conditions (21% O₂). **k,** LC–MS quantification of 5hmC levels in response to hypoxia (0.5% O₂) and glutamine-free culture conditions. **l,** Growth of cell lines cultured for 24 h under control (21% O₂, white) and hypoxic (0.5% O₂, red) conditions, as assessed using a sulforhodamine B colorimetric assay. Changes in cell density after 24 h are depicted relative to control conditions (21% O₂). **m,** IOX2-induced changes in the global 5hmC content of DNA, in *TET* mRNA expression and in hypoxia marker gene expression of five cell lines treated for 24 h with DMSO (carrier, white) or IOX2 (50 μM, blue). **n,** 5mC hydroxylation activity of nuclear lysates from MCF7 cells grown for 24 h under 21% or 0.5% O₂ (white or red). Bars represent the mean ± s.e.m. of 5 (**b, k, m**), 6 (**a, c–e**), 16 (**g–j**) or 24 (**l**) samples prepared on different days. **P* < 0.05, ***P* < 0.01, ****P* < 0.001 by *t*-test (**b, e, h–m**).



Extended Data Figure 4 | Genomic profiles of 5mC and 5hmC. Shown are results from DIP-seq of DNA from MCF7 cells cultured for 24 h under 21% or 0.5% O₂ (control and hypoxia), with examples of 5hmC-DIP-seq (top) and 5mC-DIP-seq (bottom) read depths (FPM, fragments per base pair per million fragments) at regions surrounding the transcription start site of *NSD1*, *FOXA1* and *CDKN2A*. These show 5hmC loss (FDR < 5%)

and a 5mC gain that is more subtle, perhaps because the resolution of 5mC-DIP-seq is limiting: regions rich in 5hmC tend to be poorer in 5mC⁵⁴, and thus have less substrate available for pull-down. 5mC-DIP-seq moreover captures all methylated sites, so most of the 5mC-DIP-seq signal does not derive from sites that are actively turning over 5hmC.

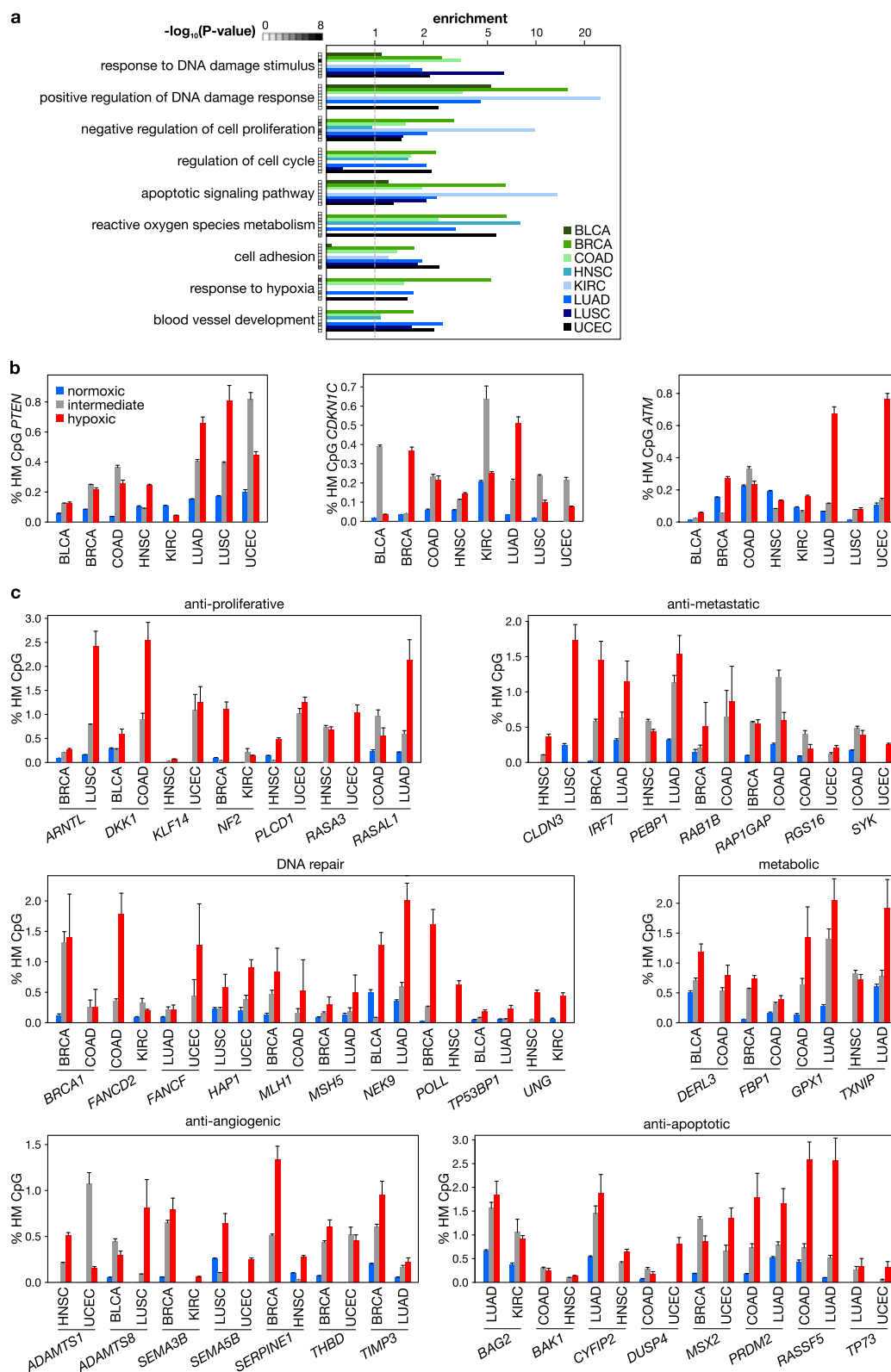


Extended Data Figure 5 | See next page for caption.

Extended Data Figure 5 | Effect of hypoxia on hypermethylation frequency in tumours.

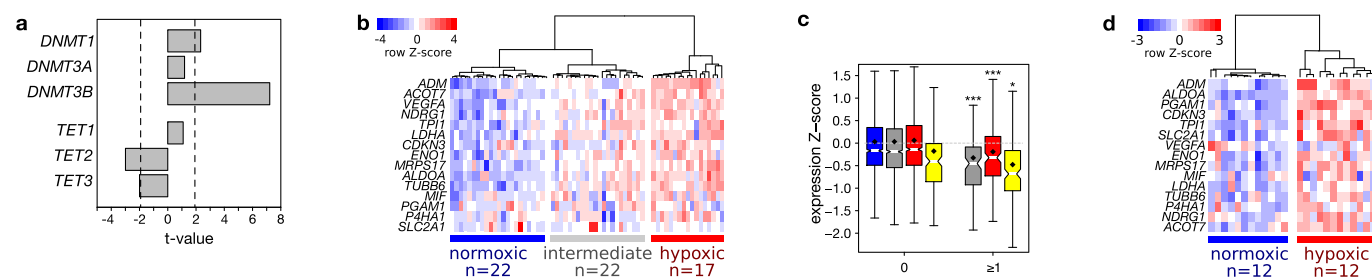
a, Immunofluorescence analysis of patient-derived tumour xenografts, stained for pimonidazole (white), 5hmC (red), DNA (propidium iodide, blue) and pan-cytokeratin (green). Shown are representative images of a breast and two endometrial tumour xenografts. The inset on the right shows box plots illustrating the signal in normoxic pimonidazole-negative nuclei (blue), and in hypoxic pimonidazole-positive nuclei (red). **b**, Hypoxia marker gene expression clusters, with the first three clusters used to define normoxic, intermediate and hypoxic tumours. **c**, Unsupervised clustering of 1,000 CpGs showing the highest average methylation increase in tumour versus corresponding normal tissues. The first three clusters were used to define tumours of low, intermediate and high hypermethylation. The colour bar above the clusters annotates each tumour as normoxic, intermediate or hypoxic, as determined in **b**. **d**, Box plots showing the relative expression (z-score) of genes in tumours in which they have either 0 or ≥ 1 hypermethylation event in their promoter, stratified into normoxic, intermediate and hypoxic tumours (blue, grey and red, respectively). Diamonds indicate mean, box plot wedges indicate $2\times$ the standard error of the median. Genes with ≥ 1 hypermethylation events in their promoters have a lower

average expression level ($P < 0.01$ for each tumour type). **e**, Fraction of genes having a promoter that is rich, intermediate or poor in CpGs, out of all gene promoters that are assessed on the 450k array, and out of all gene promoters that are frequently hypermethylated in the indicated tumour types. **f**, Fraction of 1,742 *TET* wild-type tumours and 39 *TET* mutant that are normoxic, intermediate and hypoxic. $P > 0.2$ for all comparisons. **g**, Cell proliferation marker gene⁴⁶ expression clusters, with the first two clusters used to define high-proliferative and low-proliferative tumours. **h**, hypermethylation frequencies in low- and high-proliferative tumours, with asterisks representing P values from linear models correcting for variables specified in Supplementary Table 8. **i**, Partial correlation coefficient (partial R^2) estimates of the relative contribution of tumour characteristics (annotated in TCGA) to the variance in hypermethylation observed in these tumours. Partial R^2 values were obtained from linear model estimation using ordinary least squares, and expressed as a fraction of the total variance (that is, total R^2) explained by the model when taking into account all indicated variables, as indicated between brackets under each tumour type. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$ by t -test (**a**) or by generalized linear model (**h**).



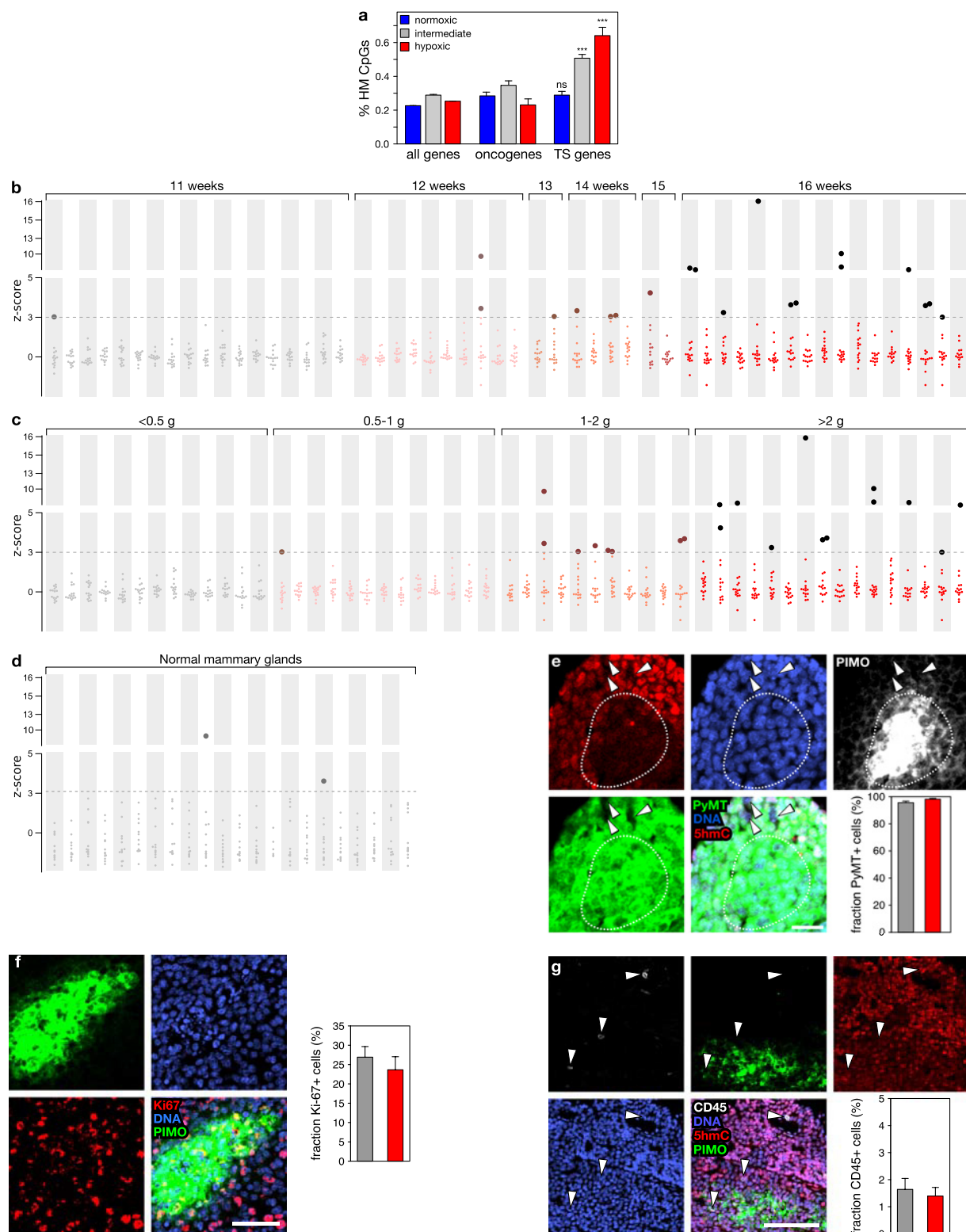
Extended Data Figure 6 | Functional annotation of genes more frequently hypermethylated in hypoxic tumours. **a**, Ontology terms enrichment analysis of genes that are more frequently hypermethylated at their gene promoters in hypoxic than normoxic tumours, for eight tumour types characterized in the TCGA pan-cancer effort. A representative set of terms is displayed, selected from terms enriched in most tumour types. *P* values as defined by the grey-scale insert. Enrichment calculated

using topGO. **b**, Selected examples of hypermethylation frequencies in the promoter of key TSGs (*PTEN*, *CDKN1C*, *ATM*) more frequently hypermethylated in normoxic than hypoxic tumours. **c**, Hypermethylation frequency in the promoter of selected genes involved in the processes indicated. *P* < 0.05 for all genes (asterisks are not displayed). Bars in **b** and **c** represent the hypermethylation frequency \pm s.e.m. *P* values in **a** by Fisher's exact test.



Extended Data Figure 7 | Effect of hypoxia on TET activity in human tumours. **a**, The t -value of correlation between hypermethylation and expression of *TET* or *DNMT* genes across 3,141 tumours of 8 tumour types (bladder, breast, colorectal, head and neck, kidney, lung adenocarcinoma, lung squamous, and uterine carcinoma) profiled in TCGA for gene expression and DNA methylation, while correcting for tumour type, hypoxia and proliferation. The dotted line represents $P < 0.05$, negative t -values represent inverse correlations. **b**, Hypoxia metagene signature applied to 63 glioblastoma multiforme tumours from TCGA. **c**, Boxplots showing the relative expression (Z-score) of genes in

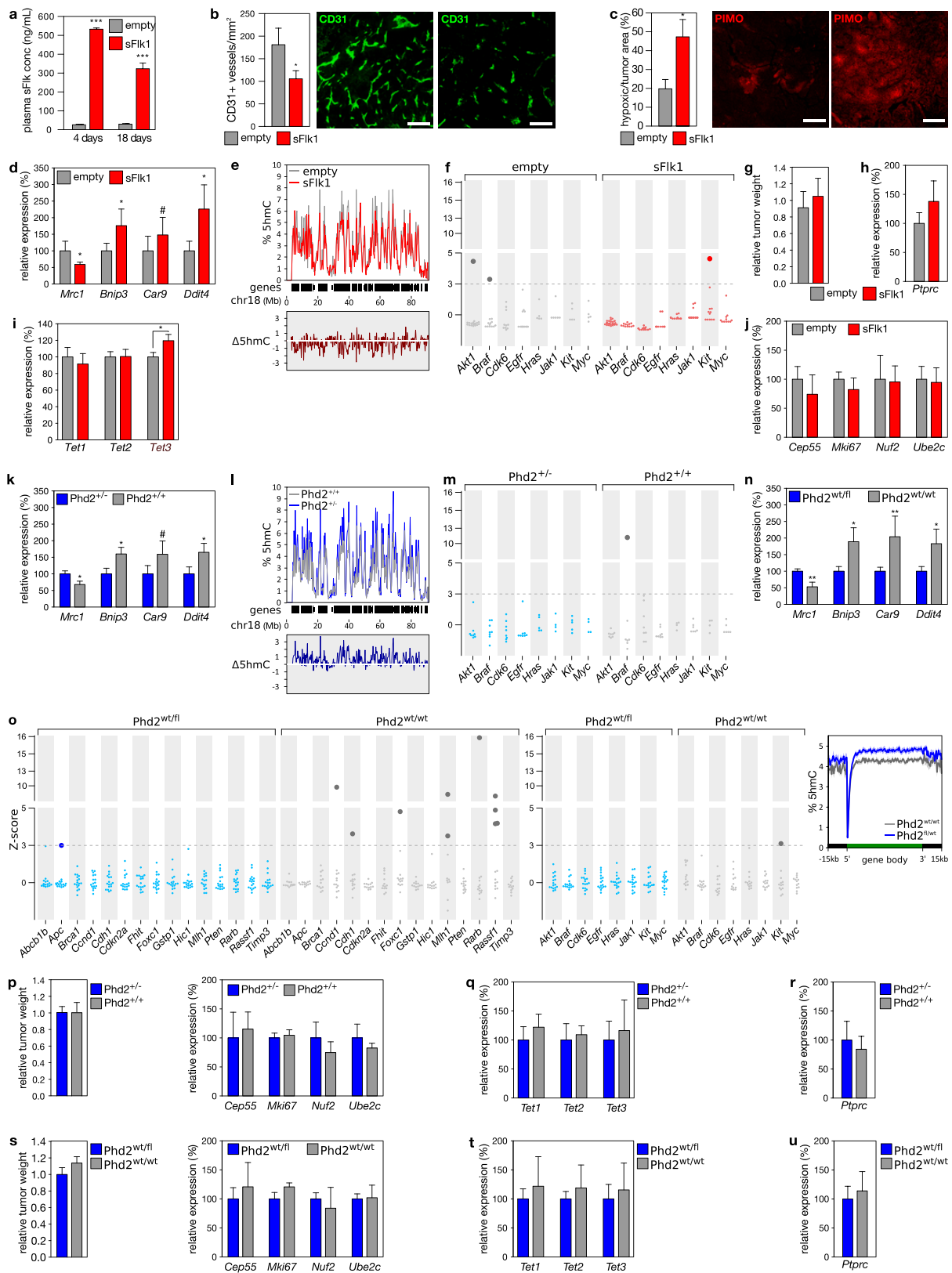
tumours in which they have either 0 or ≥ 1 hypermethylation event in their promoter, stratified into wild-type *IHD1* tumours that are normoxic ($n = 19$), intermediate ($n = 21$) and hypoxic ($n = 17$) (blue, grey and red, respectively), and *IDH1*^{R138}-mutated tumours ($n = 4$, yellow). Diamonds indicate mean, box plot wedges indicate $2 \times$ the standard error of the median. Genes with ≥ 1 hypermethylation events in their promoters have a lower average expression level. No hypermethylation events were detected in wild-type *IHD1* normoxic tumours. **d**, Hypoxia metagene signature applied to 12 normoxic and 12 hypoxic non-small-cell lung tumours. * $P < 0.05$, *** $P < 0.001$ by t -test (**c**).



Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | 5hmC, hypoxia and TSG hypermethylation in mouse breast tumours. **a**, Frequency of hypermethylation events in the promoters of all genes, all oncogenes and all TSGs as annotated²⁸, in 695 human breast tumours available through TCGA and stratified into normoxic, intermediate and hypoxic subsets. **b, c**, DNA was extracted from 53 tumours developing in MMTV-PyMT mice of the indicated ages (**c**) or weights (**d**) and sequenced to a depth of $\sim 500\times$. Plotted are z-scores of hypermethylation (y axis, exponential) for 15 TSGs, relative to the tumours from 11-week-old mice. The dotted line represents the threshold for a Bonferroni-adjusted $P < 0.05$, and bold darker dots are used for tumours displaying significantly increased hypermethylation events. **d**, DNA extracted from 20 normal mammary glands from 14-week-old mice, PCR-amplified for 15 TSGs and sequenced to a depth of $\sim 500\times$. Plotted are z-scores of hypermethylation relative to 11-week-old tumours. **e**, Staining of PyMT tumours for 5hmC (red), DNA (propidium iodide, blue), pimonidazole (white) and PyMT (green), and fraction of PyMT-positive cells in normoxic and hypoxic areas. The area outlined

corresponds to the hypoxic, pimonidazole-positive section, arrowheads point to PyMT-negative cells. Scale bar, $25\mu\text{m}$. The bar chart inset illustrates the relative number of PyMT-positive cells in normoxic and hypoxic areas (grey and red, respectively; $n = 19$). **f**, Ki67-positive cells in PyMT tumours: representative image of staining for DNA (propidium iodide, blue), Ki67 (red) and pimonidazole (green). Scalebar, $50\mu\text{m}$. The bar chart inset illustrates the quantification of Ki67-positive cells in normoxic and hypoxic areas (grey and red, respectively) across 6 tumours, analysing 3 fields of view with over 150 cells per field of view. **g**, CD45-positive cells in PyMT tumours: representative image of staining for DNA (propidium iodide, blue), 5hmC (red), pimonidazole (green) and CD45 (white). Scale bar, $100\mu\text{m}$. The bar chart inset illustrates the quantification of CD45-positive cells in normoxic and hypoxic areas (white and red, respectively) of 11 tumours, capturing on average $\sim 2,500$ nuclei per analysis. *** $P < 0.001$ in (**a**) by Fisher's exact test, significance relative to all genes.



Extended Data Figure 9 | See next page for caption.

Extended Data Figure 9 | Manipulation of tumour oxygenation in mouse breast tumours, and effects on 5hmC, TSG hypermethylation and confounders.

a, Plasma sFlk1 concentrations at the indicated times after hydrodynamic injection with an empty ($n = 7$) or sFlk1-overexpression plasmid ($n = 5$) (grey and red, respectively). **b, c** Quantification of tumour vessel number (**b**) and hypoxic areas (**c**) of tumours from transgenic MMTV-PyMT mice, hydrodynamically injected with an empty or sFlk1-overexpression plasmid, with representative images of blood vessels stained for CD31 (**b**) and hypoxic areas stained for pimonidazole adducts (**c**). Scale bar, 100 μm . **d**, Changes in RNA expression of hypoxia marker genes that are known to be downregulated (*Mrc1*) or upregulated (*Bnip3*, *Car9*, *Ddit4*) in hypoxic conditions. **e**, 5hmC levels (y axis) across mouse chromosome 18 (x axis) in 400 kb bins, with the location of RefSeq genes (middle), and differences in 5hmC levels (lower). 5hmC levels were determined using shallow TAB-seq, and chromosome 18 was selected because it has large stretches of gene deserts that illustrate the 5hmC depletion in these areas ($n = 3$). 5hmC levels decrease by $12.4\% \pm 3.5$ after sFlk1 overexpression, although technical limitations of TAB-seq (incomplete 5hmC protection or bisulfite-conversion) may partially obscure the magnitude of effects. **f**, Hypermethylation in tumours developing in 12-week-old mice receiving hydrodynamic injection with an empty ($n = 19$) or sFlk1-overexpressing plasmid ($n = 24$) 3 weeks earlier. DNA was bisulfite-converted, PCR-amplified for the indicated oncogenes, and sequenced to a depth of $\sim 500\times$. Plotted are z -scores of hypermethylation (y -axis, exponential), relative to the more normoxic tumours (that is, empty). The dotted line represents the threshold at 5% FDR, and bold darker dots the tumours displaying significantly increased hypermethylation events. **g–j**, Relative

weights of tumours from tg(MMTV-PyMT) mice, hydrodynamically injected with an empty (grey, $n = 19$) or sFlk1-overexpression plasmid (red, $n = 24$) (**g**), and corresponding RNA expression of *Ptprc* (the gene encoding CD45, $n = 5$) (**h**), of *Tet* enzymes (**i**, $n = 15$ for empty plasmid, $n = 12$ for sFlk1-overexpressing plasmid) and of cell proliferation markers (**j**, $n = 5$ for each). **k–m**, As in **d–f** but for 16-week old transgenic MMTV-PyMT mice of the indicated genotype. $n = 5$ (**k**), $n = 3$ for *Phd2*^{+/+}; $n = 4$ for *Phd2*^{fl/fl} (**l**) and $n = 9$ (**m**). **n**, As in **d**, but for 16-week-old *Tie2*-Cre;Tg(MMTV-PyMT) mice of the indicated genotypes ($n = 5$). **o**, DNA was extracted from 17 breast tumours developing in *Tie2*-Cre;*Phd2*^{fl/WT}; Tg(MMTV-PyMT) mice (blue) and 13 breast tumours developing in *Tie2*-Cre;*Phd2*^{WT/WT};Tg(MMTV-PyMT) mice (grey), all 16 weeks old. DNA was bisulfite-converted, PCR-amplified for the indicated TSGs (left) and oncogenes (middle) and sequenced to a depth of $>500\times$. Plotted are z -scores of hypermethylation (y axis, exponential), relative to the more normoxic, *Phd2*^{WT/fl}, tumours. The dotted line represents the threshold for a Bonferroni-adjusted $P < 0.05$, and bold darker dots the tumours displaying significantly increased hypermethylation events. Right, 5hmC levels \pm s.e.m. across a metagene in tumours of 16-week-old mice with the indicated genotype ($n = 3$ for *Phd2*^{fl/fl}; $n = 4$ for *Phd2*^{WT/fl}). **p–u**, Relative weights of tumours from *Phd2*^{+/+};tg(MMTV-PyMT) mice and *Phd2*^{+/+};Tg(MMTV-PyMT) mice ($n = 10$ and 9 resp.) (**p–r**) and from *Tie2*-Cre;*Phd2*^{fl/WT};Tg(MMTV-PyMT) and *Tie2*-Cre;*Phd2*^{WT/WT};Tg(MMTV-PyMT) mice ($n = 17$ and 13, respectively) (**s–u**), and the corresponding RNA expression of cell proliferation markers ($n = 5$, **p, s**), of *Tet* enzymes ($n = 5$, **q, t**) and of *Ptprc* ($n = 5$) (**r, u**). # $P < 0.10$, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ by t -test.

A combined transmission spectrum of the Earth-sized exoplanets TRAPPIST-1 b and c

Julien de Wit¹, Hannah R. Wakeford², Michaël Gillon³, Nikole K. Lewis⁴, Jeff A. Valenti⁴, Brice-Olivier Demory⁵, Adam J. Burgasser⁶, Artem Burdanov³, Laetitia Delrez³, Emmanuël Jehin³, Susan M. Lederer⁷, Didier Queloz⁵, Amaury H. M. J. Triaud⁸ & Valérie Van Grootel³

Three Earth-sized exoplanets were recently discovered close to the habitable zone^{1,2} of the nearby ultracool dwarf star TRAPPIST-1 (ref. 3). The nature of these planets has yet to be determined, as their masses remain unmeasured and no observational constraint is available for the planetary population surrounding ultracool dwarfs, of which the TRAPPIST-1 planets are the first transiting example. Theoretical predictions span the entire atmospheric range, from depleted to extended hydrogen-dominated atmospheres^{4–8}. Here we report observations of the combined transmission spectrum of the two inner planets during their simultaneous transits on 4 May 2016. The lack of features in the combined spectrum rules out cloud-free hydrogen-dominated atmospheres for each planet at $\geq 10\sigma$ levels; TRAPPIST-1 b and c are therefore unlikely to have an extended gas envelope as they occupy a region of parameter space in which high-altitude cloud/haze formation is not expected to be significant for hydrogen-dominated atmospheres⁹. Many denser atmospheres remain consistent with the featureless transmission spectrum—from a cloud-free water-vapour atmosphere to a Venus-like one.

On 4 May 2016, we observed the simultaneous transits of the Earth-sized planets TRAPPIST-1b and TRAPPIST-1c with the Hubble Space Telescope (HST). This rare event was phased with HST's visibility window of the TRAPPIST-1 system, allowing for complete monitoring of the event (Fig. 1). Observations were conducted in 'round-trip' spatial scanning mode¹⁰ using the near-infrared (1.1–1.7 μm) G141 grism on the wide-field camera 3 (WFC3) instrument (see Methods). Following standard practice, we monitored the transit event through four HST orbits, taking observations before, during and after the transit event to acquire accurate stellar baseline flux levels. We discarded the first orbit owing to differing systematics caused by the thermal settling of the telescope following target acquisition^{11–13}. The raw light curve presents primarily ramp-like systematics on the scale of HST orbit-induced instrumental settling, discussed in previous WFC3 transit studies^{11,12,14} (Fig. 1). We reduced, corrected for instrumental systematics, and analysed the data using independent methods (see Methods) that yielded consistent results. We reached an average standard deviation of normalized residuals (SDNR) of 650 parts per million (p.p.m.) per 112-second exposure (Fig. 2) on the spectrophotometric time series split in 11 channels (resolution = $\lambda/\Delta\lambda \approx 35$). Summing over the entire WFC3 spectral range, we derived a 'white' light curve with a 240-p.p.m. SDNR (Fig. 1).

We first analysed the fitting of the white-light curve for the transits of TRAPPIST-1b and TRAPPIST-1c simultaneously, while accounting for instrumental systematics. Owing to the limited phase coverage of HST observations, we fixed the system's parameters to the values provided in the discovery report³ while estimating the transit times and depths. However, we let the band-integrated limb-darkening coefficients

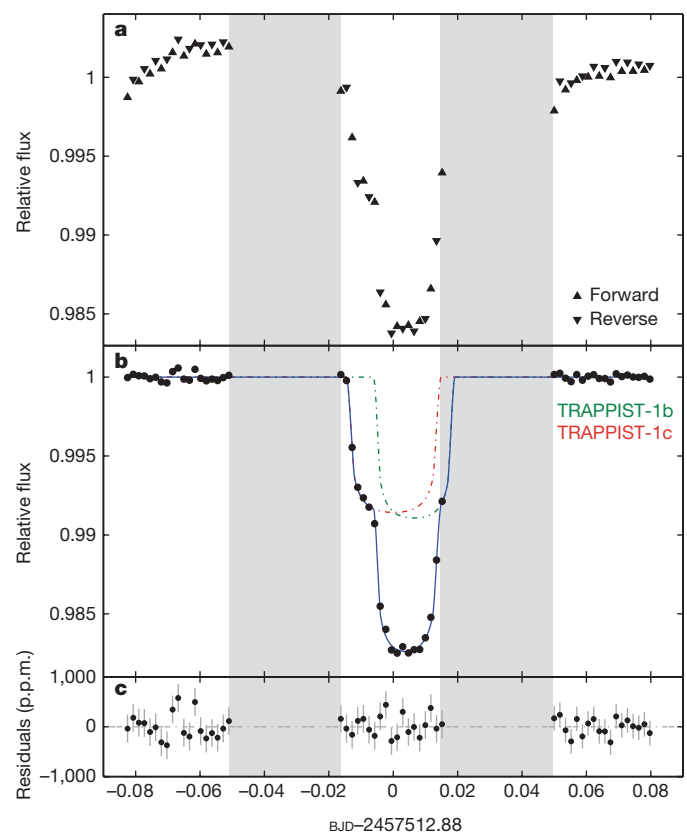


Figure 1 | Hubble/WFC3 white-light curve for the TRAPPIST-1b and TRAPPIST-1c double transit of 4 May 2016. **a**, Raw normalized white-light curve (triangles), highlighting the primary instrumental systematics (the forward/reverse flux offset and the ramp; see Methods). The shaded areas represent time windows during which no exposure was taken owing to occultation by the Earth. **b**, Normalized and systematics-corrected white-light curve (black points) and best-fit transit model (blue line). The individual contributions of TRAPPIST-1b and TRAPPIST-1c are shown in green and red, respectively. **c**, Best-fit residuals with their 1σ error bars (SDNR = 240 p.p.m.).

(LDCs) and the orbital inclinations for planets b and c (i_b and i_c , respectively) float under the control of priors, to propagate their uncertainties on the transit depth and time estimates with which they may be correlated. These priors were derived from the PHOENIX model intensity spectra¹⁵ for the LDCs (see Methods) and from the discovery report³ for the planets' orbital inclinations. We find that TRAPPIST-1c

¹Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. ²NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA. ³Institut d'Astrophysique et de Géophysique, Université de Liège, Allée du 6 Août 19C, 4000 Liège, Belgium. ⁴Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, Maryland 21218, USA. ⁵Astrophysics Group, Cavendish Laboratory, 19 J J Thomson Avenue, Cambridge CB3 0HE, UK. ⁶Center for Astrophysics and Space Science, University of California San Diego, La Jolla, California 92093, USA. ⁷NASA Johnson Space Center, 2101 NASA Parkway, Houston, Texas 77058, USA. ⁸Institute of Astronomy, Madingley Road, Cambridge CB3 0HA, UK.

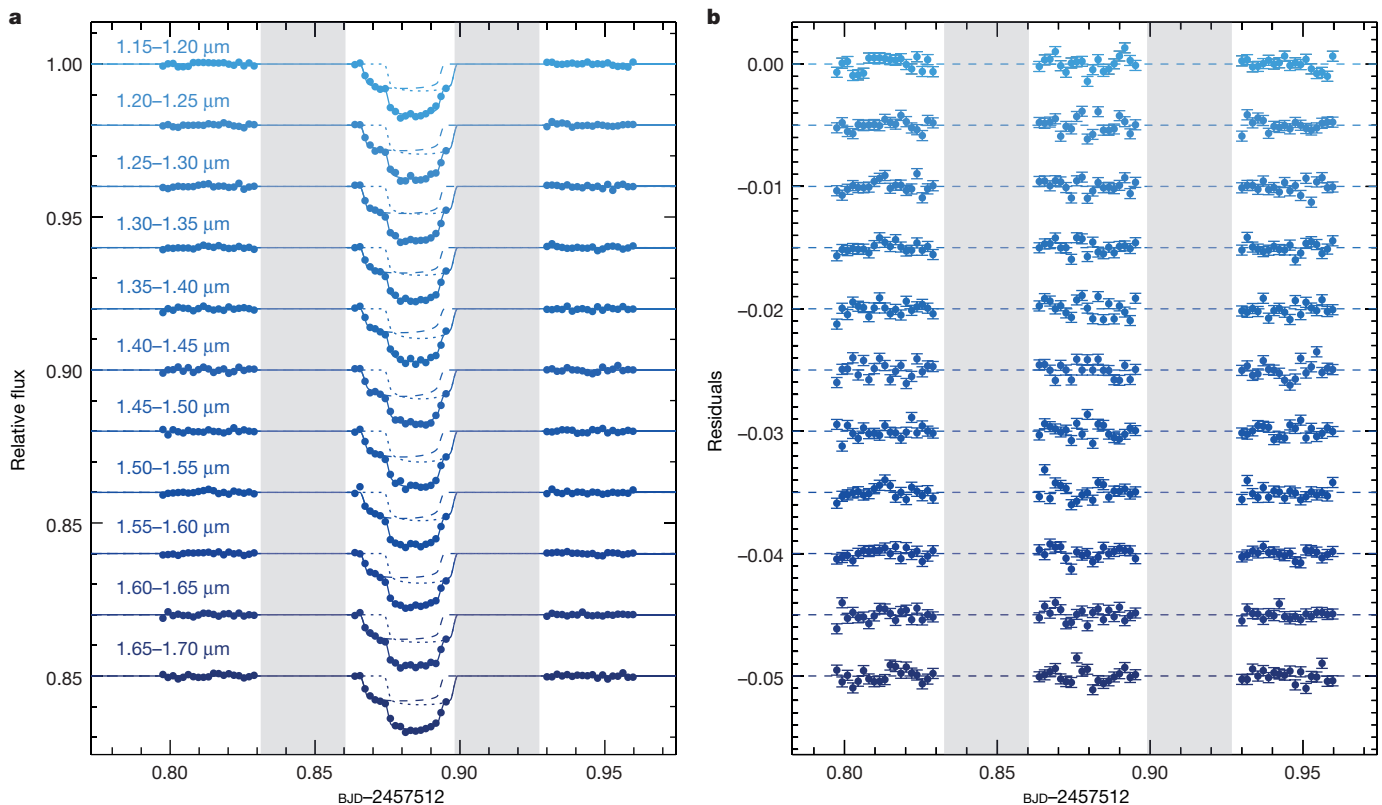


Figure 2 | Hubble/WFC3 spectrophotometry of the TRAPPIST-1b and TRAPPIST-1c double transit of 4 May 2016. a, Normalized and systematics-corrected data (points) and best-fit transit model (solid line) in 11 spectroscopic channels spread across the WFC3 band, offset for

began its transit 12 minutes before TRAPPIST-1b (transit time centres at barycentric Julian date (BJD)/barycentric coordinate time (TBD) –2457512: $T_{0,b} = 0.88646 \pm 0.00030$ and $T_{0,c} = 0.88019 \pm 0.00016$; transit durations³: $W_b = 36.12 \pm 0.46$ min and $W_c = 41.78 \pm 0.81$ min). The difference between the planets' transit duration of 5.6 ± 0.9 min implies that no planet–planet eclipse¹⁶ occurred during the observed event, given the well established orbital periods. Standard transit models¹⁷ are therefore adequate for analysing this data set. We find an orbital inclination and transit depth across the full WFC3 band of $i_b = 89.39^\circ \pm 0.32^\circ$ and $\Delta F_b = 8,015 \pm 220$ p.p.m. for TRAPPIST-1b, and $i_c = 89.58^\circ \pm 0.11^\circ$ and $\Delta F_c = 7,290 \pm 240$ p.p.m. for TRAPPIST-1c.

In the context of double-transit observations, the data primarily constrain the combined transit depths ($\Delta F_{b+c} = 15,320 \pm 160$ p.p.m.). Therefore, although the partial transit of TRAPPIST-1c—before TRAPPIST-1b begins its transit—yields some constraints on ΔF_c , it is not sufficient to completely lift the degeneracy between ΔF_b (being $\Delta F_{b+c} - \Delta F_c$) and ΔF_c . This explains the $\sim 30\%$ better precision obtained with the combined transit depth, and hence also with the combined transmission spectrum. The transit depths derived over WFC3's band are in agreement, within 2σ , with the values reported at discovery³.

We then analysed the light curves in 11 spectroscopic channels, fitting for wavelength-dependent transit depths, instrumental systematics, and stellar baseline levels (Fig. 2). We tried both quadratic and four-parameter limb-darkening relationships¹⁸ for each spectroscopic channel, because transit depth estimates may depend on the functional form used to describe limb darkening. We found, however, that our conclusions are not sensitive to which limb-darkening relationship was chosen, as long as the wavelength dependence of the LCDs is taken into account. The resulting transmission spectra are consistent with a flat line (Fig. 3).

Figure 3 shows the transit depth variations expected over the WFC3 band if TRAPPIST-1b and/or TRAPPIST-1c were harbouring

clarity. The individual contributions of TRAPPIST-1b and TRAPPIST-1c are shown with dotted and dashed lines, respectively. **b**, Best-fit residuals with their 1σ error bars (channel-averaged SDNR = 650 p.p.m.).

a cloud-free hydrogen-dominated atmosphere (red lines and circles in Fig. 3). Our transmission spectrum model¹⁹ sets atmospheric temperature to the planet's equilibrium temperature ($T_{eq,b} = 366$ K and $T_{eq,c} = 315$ K), assuming a Bond albedo of 0.3. Because the planetary masses remain unmeasured, we conservatively use a mass of $0.95M_\oplus$ and $0.85M_\oplus$ for TRAPPIST-1b and TRAPPIST-1c respectively (M_\oplus being the mass of Earth); these are the maximum masses that allow them to possess hydrogen/helium envelopes greater than 0.1% of their total masses given their radii²⁰. The precision achieved with the combined transmission spectrum (~ 350 p.p.m. per bin) is sufficient to detect the presence of a cloud-free hydrogen-dominated atmosphere via the detection of water or methane absorption features. The featureless spectra rule out a cloud-free, hydrogen-dominated atmosphere for TRAPPIST-1b and TRAPPIST-1c at the 12σ and 10σ level, respectively.

We also show in Fig. 3 alternative atmospheres for TRAPPIST-1b and TRAPPIST-1c that are consistent with the data; volatile (water)-rich atmospheres and hydrogen-dominated atmospheres with a cloud deck at 10 mbar are shown in blue and in yellow, respectively. Many alternatives for the atmospheres of TRAPPIST-1b and TRAPPIST-1c still remain. The atmospheric screening of sub-Neptune-sized exoplanets using existing observatories is a step-by-step process^{14,21,22}. As for the super-Earth-sized planet GJ 1214b (ref. 21), the first observations of TRAPPIST-1's planets with HST allow us to rule out a cloud-free hydrogen-dominated atmosphere for either planet. If the planets' atmospheres are hydrogen-dominated, then they must contain clouds or hazes that are grey absorbers between $1.1 \mu\text{m}$ and $1.7 \mu\text{m}$ at pressures less than around 10 mbar. However, theoretical investigations for hydrogen-dominated atmospheres⁹ predict that the efficiencies of haze and cloud formation at the irradiation levels of TRAPPIST-1b and TRAPPIST-1c should be dramatically reduced compared with, for example, the efficiencies for GJ 1214b (insolation ratios: $S_{GJ1214b}/S_b \approx 4$; $S_{GJ1214b}/S_c \approx 8$), leading to cloud formation at

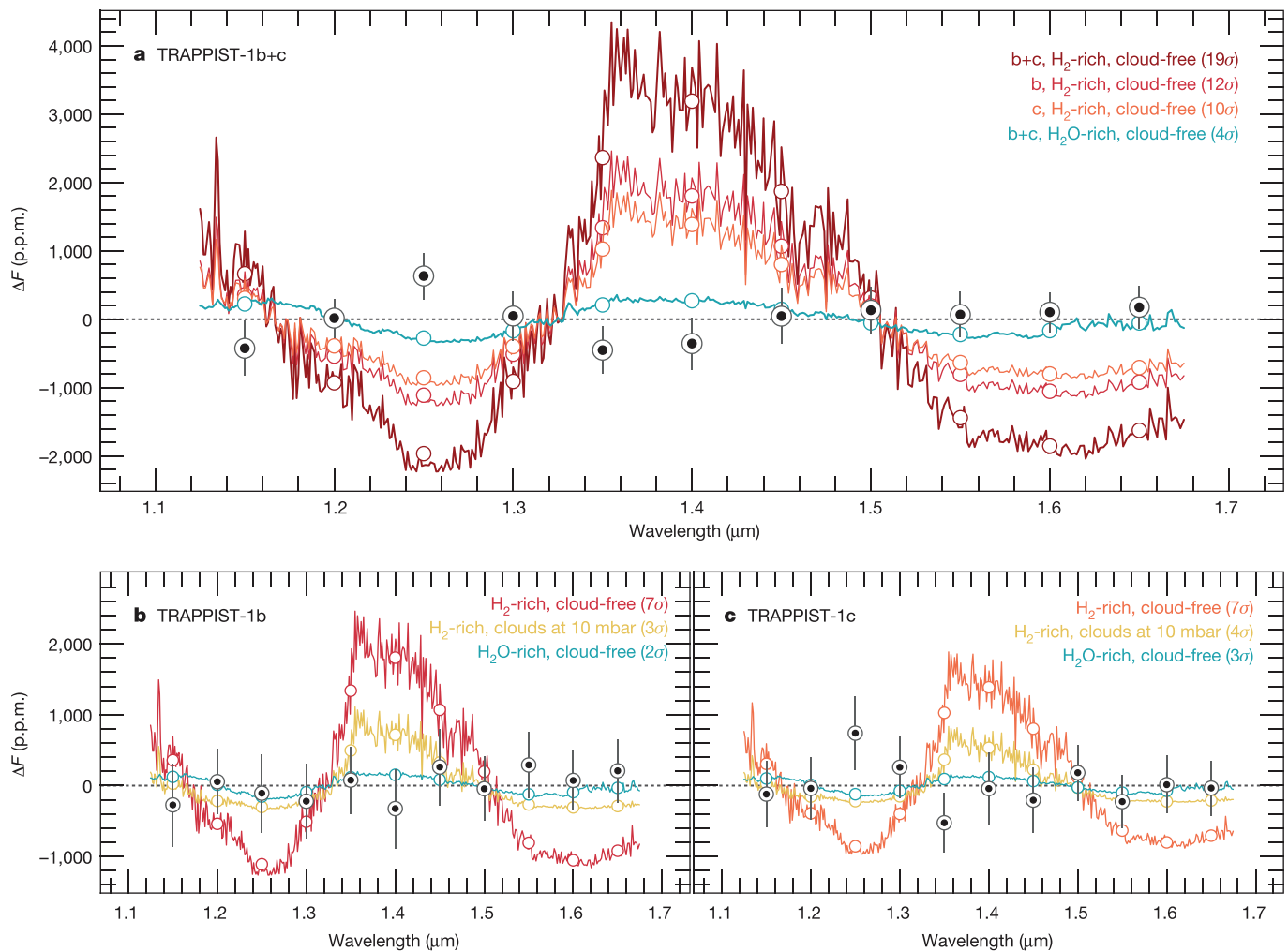


Figure 3 | Transmission spectra of TRAPPIST-1b and TRAPPIST-1c compared with models. **a–c**, Theoretical predictions of TRAPPIST-1b's transmission spectrum (**b**), TRAPPIST-1c's spectrum (**c**), and their combinations (**a**) are shown for cloud-free H_2 -dominated atmospheres (red lines and circles), H_2 -dominated atmospheres with a cloud deck at 10 mbar (yellow lines and circles), and cloud-free H_2O -dominated

atmospheres (blue lines and circles). The coloured circles show the binned theoretical models. The feature at $1.4\mu\text{m}$ arises from water absorption. The significance of the deviation of each transmission spectrum from the WFC3 measurements (black circles with 1σ error bars) is listed in parentheses in each panel.

pressure levels of 100 mbar or more, with marginal effects on their transmission spectra¹⁹. In short, hydrogen-dominated atmospheres can be considered as unlikely for TRAPPIST-1b and TRAPPIST-1c.

Planets with the sizes and equilibrium temperatures of TRAPPIST-1b and TRAPPIST-1c could possess relatively thick H_2O -, CO_2 -, N_2 - or O_2 -dominated atmospheres, or potentially tenuous atmospheres composed of a variety of chemical species^{4–8,23}. All of these denser atmospheres are consistent with our measurements. The amplitude of a planet's transmission spectrum scales directly with its atmospheric mean molecular weight, μ . The amplitude of an exoplanet's transmission spectrum can be expressed as $2R_p h_{\text{eff}}/R_*^2$, where R_p and R_* are the planetary and stellar radii, and h_{eff} is the effective atmospheric height (that is, the extent of the atmospheric annulus), which is directly proportional to the atmospheric scale height, $H = kT/\mu g$, where k is Boltzmann's constant, T is the atmospheric temperature, and g is the surface gravity. Therefore, everything else being equal, the transmission spectrum amplitude of a denser atmosphere is significantly damped compared with the one of a hydrogen-dominated atmosphere (for example, by a factor of about seven for a H_2O -dominated atmosphere). As a result, no constraint on the presence and minimum pressure level of clouds/hazes for such denser atmospheres can be inferred from our data. TRAPPIST-1b and TRAPPIST-1c could, for instance, harbour a cloud-free water-vapour atmosphere or a Venus-like atmosphere with

high-altitude hazes^{24,25}. We shall be able soon to distinguish between such atmospheres. The transmission spectrum of Venus as an exoplanet would present broad variations of about 2 p.p.m. from $0.2\mu\text{m}$ to $5\mu\text{m}$ (ref. 26), which, rescaled to the TRAPPIST-1 star, correspond to variations of about 160 p.p.m. ($2 \times R_{\text{Sun}}^2/R_{\text{TRAPPIST-1}}^2$)—currently below our errors, but eventually reachable.

Screening TRAPPIST-1's Earth-sized planets now—to distinguish progressively between their plausible atmospheric regimes, and to determine their amenability for detailed atmospheric studies—will allow the optimization of follow-up studies with the next generation of observatories. Our work highlights HST/WFC3's ability to perform the first step towards a thorough understanding of these planets' atmospheric properties.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 18 May; accepted 4 June 2016.

Published online 20 July 2016.

- Kopparapu, R. K. *et al.* Habitable zones around main-sequence stars: new estimates. *Astrophys. J.* **765**, 131 (2013).
- Zsom, A., Seager, S., de Wit, J. & Stamenkovic, V. Towards the minimum inner edge distance of the habitable zone. *Astrophys. J.* **778**, 109 (2013).

3. Gillon, M. *et al.* Temperate Earth-sized planets transiting a nearby ultracool dwarf star. *Nature* **533**, 221–224 (2016).
4. Owen, J. E. & Wu, Y. Kepler planets: a tale of evaporation. *Astrophys. J.* **775**, 105 (2013).
5. Jin, S. *et al.* Planetary population synthesis coupled with atmospheric escape: a statistical view of evaporation. *Astrophys. J.* **795**, 65 (2014).
6. Johnstone, C. P. *et al.* The evolution of stellar rotation and the hydrogen atmospheres of habitable-zone terrestrial planets. *Astrophys. J.* **815**, L12 (2015).
7. Luger, R. & Barnes, R. Extreme water loss and abiotic O₂ buildup on planets throughout the habitable zones of M dwarfs. *Astrobiology* **15**, 119–143 (2015).
8. Owen, J. E. & Mohanty, S. Habitability of terrestrial-mass planets in the HZ of M dwarfs. I. H/He-dominated atmospheres. *Mon. Not. R. Astron. Soc.* **459**, 4088–4108 (2016).
9. Morley, C. V. *et al.* Thermal emission and reflected light spectra of super Earths with flat transmission spectra. *Astrophys. J.* **815**, 110 (2015).
10. McCullough, P. & MacKenty, J. *Considerations for using spatial scans with WFC3*. Instr. Sci. Report WFC3 2012-08 (Space Telescope Science Institute, 2012).
11. Deming, D. *et al.* Infrared transmission Spectroscopy of the exoplanets HD 209458b and XO-1b using the wide field camera-3 on the Hubble Space Telescope. *Astrophys. J.* **774**, 95 (2013).
12. Wakeford, H. R., Sing, D. K., Evans, T., Deming, D. & Mandell, A. Marginalizing instrument systematics in HST WFC3 transit light curves. *Astrophys. J.* **819**, 10 (2016).
13. Sing, D. K. *et al.* A continuum from clear to cloudy hot-Jupiter exoplanets without primordial water depletion. *Nature* **529**, 59–62 (2016).
14. Kreidberg, L. *et al.* Clouds in the atmosphere of the super-Earth exoplanet GJ1214b. *Nature* **505**, 69–72 (2014).
15. Husser, T.-O. *et al.* A new extensive library of PHOENIX stellar atmospheres and synthetic spectra. *Astron. Astrophys.* **553**, A6 (2013).
16. Hirano, T. *et al.* Planet-planet eclipse and the Rossiter-McLaughlin effect of a multiple transiting system: joint analysis of the Subaru spectroscopy and the Kepler photometry. *Astrophys. J.* **759**, L36 (2012).
17. Mandel, K. & Agol, E. Analytic light curves for planetary transit searches. *Astrophys. J.* **580**, L171–L175 (2002).
18. Sing, D. K. Stellar limb-darkening coefficients for CoRoT and Kepler. *Astron. Astrophys.* **510**, A21 (2010).
19. de Wit, J. & Seager, S. Constraining exoplanet mass from transmission spectroscopy. *Science* **342**, 1473–1477 (2013).
20. Howe, A. R., Burrows, A. & Verne, W. Mass-radius relations and core-envelope decompositions of super-Earths and sub-Neptunes. *Astrophys. J.* **787**, 173 (2014).
21. Bean, J. L., Miller-Ricci Kempton, E. & Homeier, D. A ground-based transmission spectrum of the super-Earth exoplanet GJ 1214b. *Nature* **468**, 669–672 (2010).
22. Berta, Z. K. *et al.* The flat transmission spectrum of the super-Earth GJ1214b from wide field camera 3 on the Hubble Space Telescope. *Astrophys. J.* **747**, 35 (2012).
23. Leconte, J., Forget, F. & Lammer, H. On the (anticipated) diversity of terrestrial planet atmospheres. *Exp. Astron.* **40**, 449–467 (2015).
24. Tellmann, S., Pätzold, M., Häusler, B., Bird, M. K. & Tyler, G. L. Structure of the Venus neutral atmosphere as observed by the Radio Science experiment VeRa on Venus Express. *J. Geophys. Res. Planets* **114**, E00B36 (2009).
25. Wilquet, V. *et al.* Preliminary characterization of the upper haze by SPICAV/SOIR solar occultation in UV to mid-IR onboard Venus Express. *J. Geophys. Res. Planets* **114**, E00B42 (2009).
26. Ehrenreich, D. *et al.* Transmission spectrum of Venus as a transiting exoplanet. *Astron. Astrophys.* **537**, L2 (2012).

Acknowledgements This work is based on observations made with the NASA/ESA Hubble Space Telescope that were obtained at the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc. These observations are associated with program HST-GO-14500 (principal investigator J.d.W.), support for which was provided by NASA through a grant from the Space Telescope Science Institute. The research leading to our results was funded in part by the European Research Council (ERC) under the FP/2007-2013 ERC grant 336480, and through an Action de Recherche Concertée (ARC) grant financed by the Wallonia-Brussels Federation. H.R.W. acknowledges support through an appointment to the NASA Postdoctoral Program at Goddard Space Flight Center, administered by the Universities Space Research Association through a contract with NASA. M.G. is Research Associate at the Belgian Fonds (National) de la Recherche Scientifique (FRS-FNRS). L.D. acknowledges support of the Fund for Research Training in Industry and Agriculture of the FRS-FNRS. We thank D. Taylor, S. Deustua, P. McCullough, and N. Reid for their assistance in planning and executing our observations. We are also grateful for discussions with Z. Berta-Thompson and Pierre Magain about this study and manuscript. We thank the ATLAS and PHOENIX teams for providing stellar models.

Author Contributions J.d.W. and H.R.W. led the data reduction and analysis, with the support of M.G., N.K.L. and B.-O.D. J.d.W., H.R.W., and N.K.L. led the data interpretation, with the support of M.G. and J.A.V. J.A.V. provided the limb-darkening coefficients and further insights into TRAPPIST-1's properties and emission together with A.J.B. Every author contributed to writing both the manuscript and the HST proposal behind these observations.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.d.W. (jdw@mit.edu).

Reviewer Information *Nature* thanks D. Ehrenreich and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

HST WFC3 observations. We observed the transit of TRAPPIST-1c followed 12 minutes later by the transit of TRAPPIST-1b on 4 May 2016. Observations were conducted using the HST/WFC3 infrared G141 grism (1.1–1.7 μm) in round-trip scanning mode¹⁰. Using the round-trip scanning mode involves exposing the telescope during an initial forward slew in the cross-dispersion direction, and exposing during an equivalent slew in the reverse direction (details on the trade-offs behind round-trip scanning are below). Scans were conducted at a rate of ~ 0.236 pixels per second, with a final spatial scan covering ~ 26.4 pixels in the cross-dispersion direction on the detector.

We use the IMA output files from the CalWF3 pipeline, which have been calibrated using flat fields and bias subtraction. We applied two different extraction techniques which lead to the same conclusions. The first technique extracts the flux for TRAPPIST-1 from each exposure by taking the difference between successive non-destructive reads. A top-hat filter²⁷ is then applied around the target spectrum, measured ± 18 pixels from the centre of the TRAPPIST-1 scan, and sets all external pixels to zero. Next, the images are reconstructed by adding the individual reads per exposure back together. Using the reconstructed images, we extracted the spectra with an aperture of 31 pixels around the computed centring profile for both forward and reverse scan observations. The centring profile is calculated on the basis of the pixel flux boundaries of each exposure, which was found to be fully consistent across the spectrum for both scan directions.

The second technique uses the final science image for each exposure and determines for each frame the centroid of the spectrum in a box 28 pixels by 136 pixels, which corresponds to the dimensions of the irradiated region of WFC3's detector for our present observations. It then extracts the flux for 120 apertures of sizes ranging along the dispersion direction from 24 pixels to 38 pixels (with 1-pixel increments), and along the cross-dispersion direction from 120 pixels to 176 pixels (with 8-pixel increments)—we found the SDNR to be mostly insensitive to the aperture size along the dispersion direction. The best aperture was selected via minimization of the SDNR of the white-light-curve best fit, which is minimum for an aperture of 32 pixels by 157 pixels.

Both techniques subtract the background for each frame by selecting a region well away from the target spectrum, calculating the median flux, and cleaning cosmic-ray detections with a customized procedure²⁸. Our observations present three cosmic-ray detections that were not flagged by the CalWF3 pipeline. The exposure times were converted from Julian date in universal time (JDUT) to the barycentric Julian date in the barycentric dynamical time (BJD_{TDB}) system²⁹. Both extraction methods result in the same relative flux measurements from the star and SDNR (~ 240 p.p.m. in the white-light curve), as the build-up of flux over successive reads is stable.

We elected to obtain our observations using the round-trip scan mode in order to increase the integration efficiency compared with the standard forward scan mode. We note that, owing to slight differences in scan length/position and to the way in which the detector is read out (that is, if the direction of the scan is in the same direction as the column readout, then the integration time will be marginally longer than if the reverse were true¹⁰), round-trip scan mode results in measurable differences in the total flux of the forward scan exposures compared with the reverse scan exposures. This effect has been seen for previous WFC3 observations^{14,30} in round-trip mode, and has been corrected for in two main ways.

The first method involves splitting the data into two sets, one for forward scan exposures and one for reverse scan exposures, effectively halving the number of exposures per light curve, but doubling the number of light curves obtained. Each of these data sets is then analysed separately and the results combined at the end¹⁴. The second method uses the median of each scan direction to normalize the two light curves, which are then recombined and normalized before the light-curve analysis to obtain the transit parameters³⁰. In the TRAPPIST-1 data, we measure a $\sim 0.1\%$ difference in flux level between the two scans. Because of the limited phase coverage of the combined transits, to retain the most information about the combined and separate effects of each planet (the transit of TRAPPIST-1c followed by that of TRAPPIST-1b), we cannot apply the first method. However, by applying the second method we found significant remaining structure in the residuals, suggesting that the correction is only partial. Previous observations using the round-trip scan³⁰ show that the offset between the light curves obtained with each scan varies significantly from orbit to orbit, suggesting that correcting via a median combine across visits is not optimal. In addition, the total flux is affected asymmetrically by other instrumental systematics—for example, the detector ramp consistently yields a first measurement in the forward direction that is significantly lower than average—thus biasing the median combine. Therefore, we corrected for the flux offset induced by the round-trip scan mode on the basis of the offset in the residuals for each HST orbit individually. To do so, we estimate in our forward model the ‘intermediate residuals’, based on the data corrected for the transit model

and the instrumental systematics. For each orbit, we estimate the mean of these residuals for each scan direction (m_f and m_r for the mean of the residuals of the forward-scan exposures and the reverse-scan exposures, respectively). The ratio of the fluxes measured in reverse-scan exposures to the shared baseline level is $1 + m_r$; the ratio is $1 + m_f$ for forward-scan exposures. We therefore correct for their offsets by dividing each set of exposures by their respective ratio.

HST WFC3 white-light curve and spectroscopy. We first analysed the white-light curve by summing the flux across all wavelengths. We fitted the transits of TRAPPIST-1b and TRAPPIST-1c by using the transit model of ref. 17, while correcting for instrumental systematics. We followed the standard procedure for analysing HST/WFC3 data by fixing the planets’ orbital configurations—all but the orbital inclinations, which are currently poorly constrained for TRAPPIST-1’s planets—to the ones reported in the discovery report³, while determining the transit times and depths. We used priors on the band-integrated limb-darkening coefficients (LDCs) derived from the PHOENIX model intensity spectra¹⁵, and on the planets’ orbital inclinations—these parameters being potentially correlated with the transit depth estimates—to adequately account for our present state of knowledge on TRAPPIST-1. We used different analysis methods to confirm the robustness of our conclusions.

The first method uses a least-squares minimization fitting (L–M) implementation¹² to investigate a large sample of systematic models—which include corrections in time, HST orbital phase, and positional shifts in wavelength on the detector—and marginalize over all possible combinations to obtain the transit parameters. The L–M implementation fits the light curves for each systematic model and approximates the evidence-based weight of each systematic model using the Akaike information criterion³¹. It does so while keeping the LDCs fixed to the best estimates presented below, and the orbital inclinations fixed to the estimates from ref. 3. The highest weighted systematic models include linear corrections in time, as well as linear corrections in HST orbital phase or in the shift in wavelength position over the course of the visit. Therefore, using marginalization across a grid of stochastic models allows us to account for all tested combinations of systematics and to obtain robust transit depths for both planets, separately and in combination. For this data set, the evidence-based weight approximated for each of the systematic models applied to the data indicates that all of the systematic models fit equally well to the data, and that no one systematic model contributes to the majority of the corrections required to obtain the precision presented (Extended Data Fig. 1). In other words, instrumental systematics affect our observations only marginally. We carried out independent analyses of the data by using adaptive Markov chain Monte Carlo (MCMC) implementations^{32,33}. For each HST light curve, the transit models¹⁷ of TRAPPIST-1b and TRAPPIST-1c are multiplied by baseline models that account for the visit-long trend observed in WFC3 light curves, WFC3’s ramp, and the ‘HST breathing’ effect¹². For these analyses, priors are used for the LDCs and the orbital inclinations. We find that the visit-long trend is adequately accounted for with a linear function of time, the ramp with a single exponential in time, and the breathing with a second-order polynomial in HST’s orbital phase. More-complex baseline models were tested and gave consistent results, as revealed by the marginalization study.

We calculated the transmission spectrum by fitting the transit depth of TRAPPIST-1b and TRAPPIST-1c simultaneously in each spectroscopic light curve. We divided the spectral range between 1.15 μm and 1.7 μm into 11 equal bins of $\Delta\lambda = 0.05 \mu\text{m}$. We applied again the two techniques described above to analyse each spectroscopic light curve, resulting in the combined and independent transmission spectra of TRAPPIST-1b and TRAPPIST-1c. An L–M implementation¹² and the adaptive MCMC implementations produced consistent results for each stage of the analysis.

Limb-darkening coefficients. We determined limb-darkening coefficients by fitting theoretical specific intensity spectra (I) downloaded from the Göttingen spectral library (http://phoenix.astro.physik.uni-goettingen.de/?page_id=73), which is described in ref. 15. The intensity spectra are provided on a wavelength grid with 1-Å cadence for 78 μ values, where μ is the cosine of the angle between an outward radial vector and the direction towards the observer at a point on the stellar surface. We integrated I over one broad and 11 narrow wavelength intervals, used in our analysis of the transit light curve. We divided I for each wavelength interval by I_c , the value of I at the centre of the stellar disc (where $\mu = 1$).

Because the PHOENIX code calculates specific intensity spectra in spherical geometry, the PHOENIX μ grid extends above the stellar limb relevant to exoplanet transit calculations. When fitting limb-darkening functions, PHOENIX μ values should be scaled to yield $\mu' = 0$ at the stellar radius³⁴. We define $\mu' = (\mu - \mu_0)/(1 - \mu_0)$, where $I/I_c = 0.01$ at $\mu = \mu_0$. The value of μ_0 is a function of wavelength. We then fitted two commonly used functional forms for limb darkening¹⁸:

$$I/I_c = 1 - a(1 - \mu') - b(1 - \mu')^2$$

and

$$I/I_c = 1 - c_1(1 - (\mu')^{1/2}) - c_2(1 - \mu') - c_3(1 - (\mu')^{3/2}) - c_4(1 - \mu')^2$$

When fitting, we ignored points with $\mu' < 0.05$.

Extended Data Fig. 2 shows the limb-darkening fits for the 12 wavelength intervals in our transit light curve analysis. We calculated fits for four stellar models with effective temperatures of 2,500 K and 2,600 K and logarithmic surface gravities of 5.0 and 5.5. We then linearly interpolated the limb-darkening coefficients to an effective temperature of 2,550 K and gravity 5.22, appropriate for TRAPPIST-1 (ref. 3).

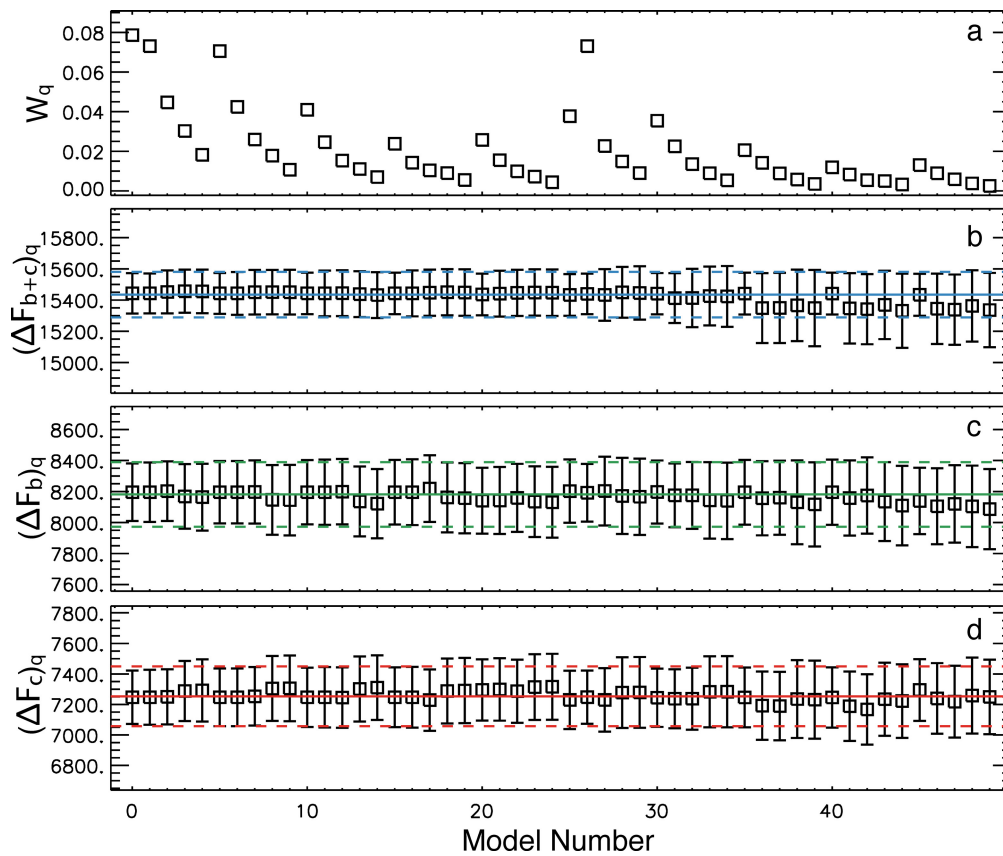
Transmission spectrum models. We simulated the theoretical spectra for TRAPPIST-1b and TRAPPIST-1c using the model introduced in ref. 19. We used atmospheric temperatures equal to the planets' equilibrium temperatures assuming a Bond albedo of 0.3 (these temperatures being 366 K for TRAPPIST-1b and 315 K for TRAPPIST-1c). The use of isothermal temperature profiles set at the equilibrium temperatures is conservative, as it does not account for possible additional heat sources or temperature inversion and results in a possible underevaluation of the atmospheric scale height. Our assumption regarding the temperature profiles does not affect our conclusion; variations of 50 K (that is, $\sim 15\%$) in the atmospheric temperature modify the amplitude of the transmission spectra by up to $\sim 15\%$, because at first order their amplitudes scale with the temperature. The planetary masses being unconstrained, we conservatively use a mass of $0.95M_\oplus$ and $0.85M_\oplus$ for TRAPPIST-1b and TRAPPIST-1c respectively—the maximum masses that would allow them to possess hydrogen/helium envelopes greater than 0.1% of their total masses given their radii²⁰. We use the atmospheric compositions of the 'mini-Neptune' and 'Halley world' models introduced in ref. 35 to simulate the hydrogen-dominated and water-dominated atmospheres, respectively. We simulated the effect of optically thick cloud or haze at a given pressure level by setting to zero the transmittance of atmospheric layers with a higher pressure.

The feature at $1.4\mu\text{m}$ arises from water absorption; the feature at $1.15\mu\text{m}$ for the water-dominated atmosphere arises from methane absorption. We compared the transmission spectra, allowing for a vertical offset to account for our *a priori* ignorance of the optically thick radius by setting the mean of each spectrum to zero. The significance of the deviation of each transmission spectrum from the WFC3 measurements is shown in Fig. 3. Significance levels less than 3σ mean that the data are consistent with that model within the reported errors.

We rule out the presence of a cloud-free hydrogen-dominated atmosphere for either planet at the 10σ level through the combined transmission spectrum (and at a lesser 7σ level through their individual spectra). The measurements are consistent with volatile (for example, water)-rich atmospheres or hydrogen-dominated atmospheres with optically thick clouds or hazes located at larger pressures than 10 mbar.

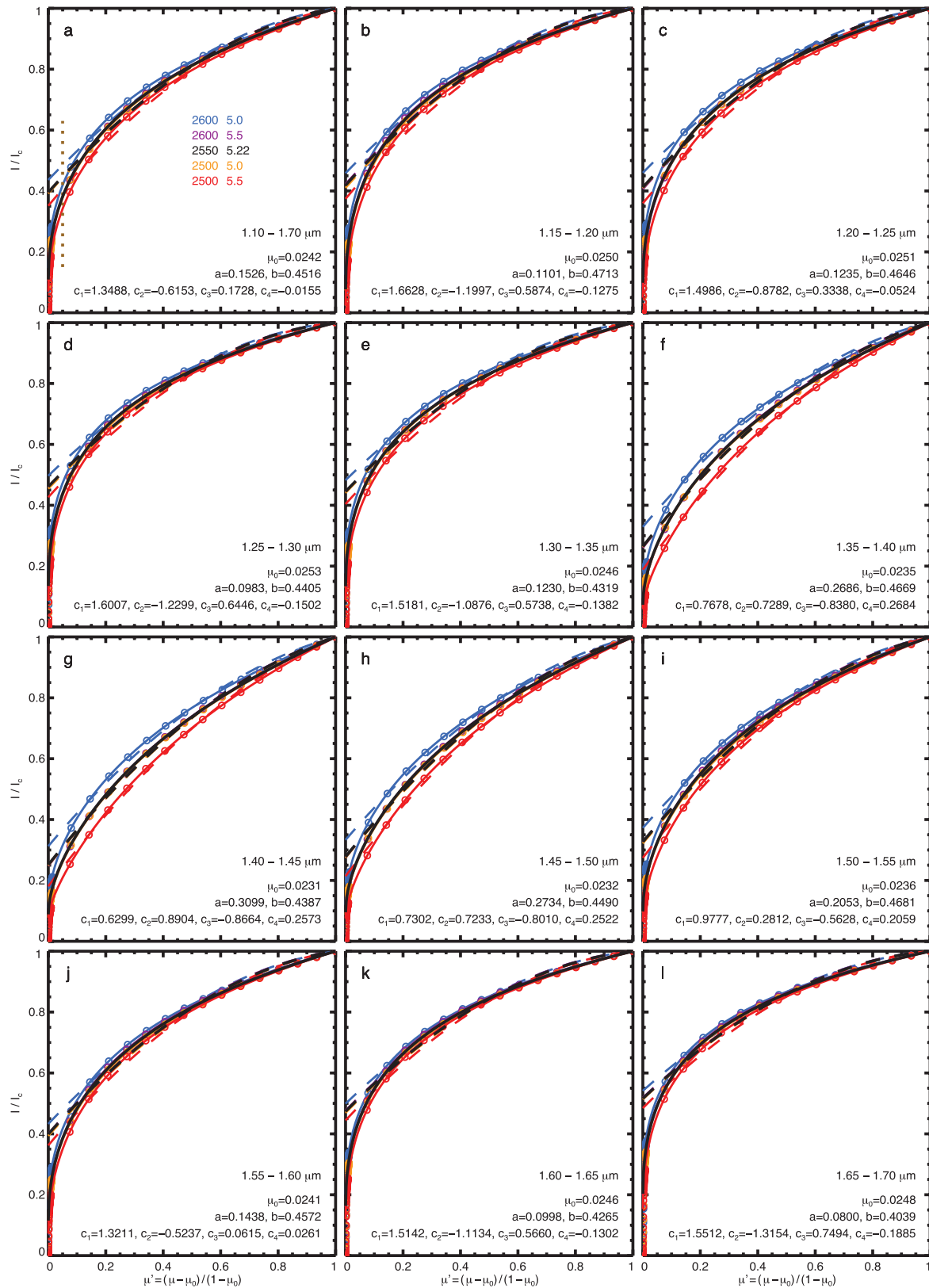
Code availability. Conversion of the UT times for the photometric measurements to the BJD_{TDB} system was carried out using the online program created by J. Eastman and distributed at <http://astroutils.astronomy.ohio-state.edu/time/utc2bjd.html>. We have opted not to make available the codes used for data extraction as they are an important part of the researchers' toolkits. For the same reason, we have opted not to make available all but one of the codes used for data analysis. The MCMC software used by M.G. to analyse independently the photometric data is a custom Fortran 90 code that can be obtained upon request. The custom IDL code used to determine limb-darkening coefficients can be obtained upon request.

27. Evans, T. M. *et al.* Detection of H₂O and evidence for TiO/VO in an ultra-hot exoplanet atmosphere. *Astrophys. J.* **822**, L4 (2016).
28. Huitson, C. M. *et al.* An HST optical-to-near-IR transmission spectrum of the hot Jupiter WASP-19b: detection of atmospheric water and likely absence of TiO. *Mon. Not. R. Astron. Soc.* **434**, 3252–3274 (2013).
29. Eastman, J., Siverd, R. & Gaudi, B. S. Achieving better than 1 minute accuracy in the heliocentric and barycentric Julian dates. *Publ. Astron. Soc. Pacif.* **122**, 935–946 (2010).
30. Knutson, H. A. *et al.* Hubble Space Telescope near-IR transmission spectroscopy of the super-Earth HD 97658b. *Astrophys. J.* **794**, 155 (2014).
31. Gibson, N. P. Reliable inference of exoplanet light-curve parameters using deterministic and stochastic systematics models. *Mon. Not. R. Astron. Soc.* **445**, 3401–3414 (2014).
32. Gillon, M. *et al.* The TRAPPIST survey of southern transiting planets. I. Thirty eclipses of the ultra-short period planet WASP-43 b. *Astron. Astrophys.* **542**, A4 (2012).
33. de Wit, J. *et al.* Direct measure of radiative and dynamical properties of an exoplanet atmosphere. *Astrophys. J.* **820**, L33 (2016).
34. Espinoza, N. & Jordán, A. Limb darkening and exoplanets: testing stellar model atmospheres and identifying biases in transit parameters. *Mon. Not. R. Astron. Soc.* **450**, 1879–1899 (2015).
35. Benneke, B. & Seager, S. Atmospheric retrieval for super-Earths: uniquely constraining the atmospheric composition with transmission spectroscopy. *Astrophys. J.* **753**, 100 (2012).



Extended Data Figure 1 | Marginal effects of instrumental systematics on transit depth estimates. **a**, Evidence-based weight, W_q , for each systematic model¹² applied to the white-light curve. **b**, Combined transit depth estimate $(\Delta F_{b+c})_q$ obtained by correcting the data, using each systematic model. **c**, **d**, Individual transit depth estimates for

TRAPPIST-1b and TRAPPIST-1c, ΔF_b and ΔF_c . The horizontal lines indicate the final marginalized measurements and associated uncertainties. The scale of the values here indicates that all of the systematic models fit equally well to the data.



Extended Data Figure 2 | TRAPPIST-1's limb darkening. Stellar limb-darkening relationships for TRAPPIST-1 (black curves) and four stellar models (coloured curves) that bracket the effective temperature and surface gravity of TRAPPIST-1 (shown in coloured and black numbers in **a**; temperature is in K and surface gravity is expressed in $\log(g)$). The circles are theoretical¹⁵ specific intensities (I) relative to disc centre (I_c) as a function of μ' (the cosine of the angle between an outward radial vector

and the direction towards the observer). We fitted I/I_c averaged over the indicated wavelength intervals to determine the quadratic (dashed curves) and four-parameter (solid curves) limb-darkening coefficients. **a**, Stellar limb-darkening relationship integrated over WFC3's spectral band. **b–l**, Stellar limb-darkening relationship over the 11 spectral channels used here.

Aggregate dust particles at comet 67P/Churyumov–Gerasimenko

Mark S. Bentley^{1*}, Roland Schmied^{1*}, Thurid Mannel^{1,2*}, Klaus Torkar¹, Harald Jeszenszky¹, Jens Romstedt³, Anny-Chantal Levasseur-Regourd⁴, Iris Weber⁵, Elmar K. Jessberger⁵, Pascale Ehrenfreund^{6,7}, Christian Koeberl^{8,9} & Ove Havnes¹⁰

Comets are thought to preserve almost pristine dust particles, thus providing a unique sample of the properties of the early solar nebula. The microscopic properties of this dust played a key part in particle aggregation during the formation of the Solar System^{1,2}. Cometary dust was previously considered to comprise irregular, fluffy agglomerates on the basis of interpretations of remote observations in the visible and infrared^{3–6} and the study of chondritic porous interplanetary dust particles⁷ that were thought, but not proved, to originate in comets. Although the dust returned by an earlier mission⁸ has provided detailed mineralogy of particles from comet 81P/Wild, the fine-grained aggregate component was strongly modified during collection⁹. Here we report *in situ* measurements of dust particles at comet 67P/Churyumov–Gerasimenko. The particles are aggregates of smaller, elongated grains, with structures at distinct sizes indicating hierarchical aggregation. Topographic images of selected dust particles with sizes of one micrometre to a few tens of micrometres show a variety of morphologies, including compact single grains and large porous aggregate particles, similar to chondritic porous interplanetary dust particles. The measured grain elongations are similar to the value inferred for interstellar dust and support the idea that such grains could represent a fraction of the building blocks of comets. In the subsequent growth phase, hierarchical agglomeration could be a dominant process¹⁰ and would produce aggregates that stick more easily at higher masses and velocities than homogeneous dust particles¹¹. The presence of hierarchical dust aggregates in the near-surface of the nucleus of comet 67P also provides a mechanism for lowering the tensile strength of the dust layer and aiding dust release¹².

MIDAS, the Micro-Imaging Dust Analysis System^{13,14}, is the first space-borne atomic force microscope (AFM) and a unique instrument designed to measure the size, shape, texture and microstructure of cometary dust. Flying on the Rosetta spacecraft, it collects dust on sticky targets during passive exposures and images its three-dimensional topography with an unprecedented nanometre to micrometre resolution¹³.

Cometary dust was first collected in mid-November 2014. Here, we focus on particles collected from then until the end of February 2015. The collected particles cover a range of sizes from tens of micrometres down to a few hundred nanometres, and have various morphologies, from single grains to aggregate particles with different packing densities. Five examples are presented here.

Figure 1 shows topographic images (height fields) of three particles (A, B and C). We refer to particles A and C as compact, because their sub-units (hereafter grains) are tightly packed; particle B appears to be a homogeneous grain. The next example (D) is also a compact particle, scanned with a higher lateral resolution of 80 nm (Fig. 2)—a factor four better than the previous scan. The final particle (E), presented in Fig. 3,

is best described as a loosely packed, ‘fluffy’ aggregate comprising many grains. Detailed collection times and geometries for all particles can be found in Extended Data Figs 1–3.

Aided by the three-dimensional nature of the data, individual grains can be identified, as shown in Figs 1b, 2b and 3b. The properties of these particles and their grains are summarized in Table 1 for particles A–D and in Fig. 3d for particle E. Because particle E extends beyond the edge of the scanned area, only lower limits for its dimensions can be given. All further calculations and discussion refer to only this visible region.

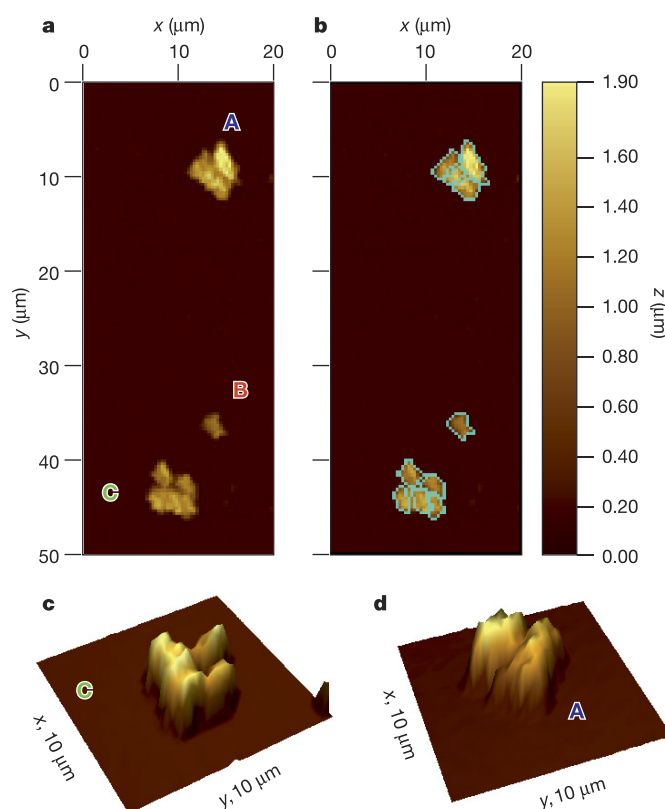


Figure 1 | AFM topographic images of particles A, B and C and their sub-units. **a**, A 20 μm × 50 μm overview image with a pixel resolution of 312 nm and the colour scale representing the height, *z*. **b**, As in **a**, but with particle B and the sub-units of particles A and C outlined in cyan. **c**, **d**, 10 μm × 10 μm three-dimensional (rotated) images of particles C and A with two-times height exaggeration to aid visualization.

¹Space Research Institute, Austrian Academy of Sciences, Schmiedlstrasse 6, 8042 Graz, Austria. ²Physics Institute, University of Graz, Universitätsplatz 5, 8010 Graz, Austria. ³European Space Research and Technology Centre, Future Missions Office (SREF), Noordwijk, The Netherlands. ⁴UPMC (Sorbonne Université); CNRS/INSU; LATMOS-IPSL, BC 102, 4 place Jussieu, 75005 Paris, France. ⁵Institut für Planetologie, Universität Münster, Wilhelm-Klemm-Strasse 10, 48149 Münster, Germany. ⁶Leiden Observatory, Postbus 9513, 2300 RA Leiden, The Netherlands. ⁷Space Policy Institute, George Washington University, 20052 Washington DC, USA. ⁸Department of Lithospheric Research, University of Vienna, Althanstrasse 14, 1090 Vienna, Austria. ⁹Natural History Museum, Burgring 7, 1010 Vienna, Austria. ¹⁰Department of Physics and Technology, UiT The Arctic University of Norway, N-9037 Tromsø, Norway.

*These authors contributed equally to this work.

Compact particles A and C are both approximately $5.6\mu\text{m}$ in effective diameter (hereafter size; see Methods) and are built from grains in the size range $1.93^{+0.10}_{-1.22}\mu\text{m}$ to $3.31^{+0.06}_{-1.23}\mu\text{m}$ (where the errors are given as the linear addition of the 1σ statistical uncertainty and the systematic uncertainty; see Methods). The compact grain B is $2.76^{+0.07}_{-0.61}\mu\text{m}$ in size, comparable to the dust grains of particles A and C. In fact, the topographic image suggests that grain B was originally part of particle C, but detached on impact with the target. Particle D is $1.09^{+0.01}_{-0.25}\mu\text{m}$ in size, again similar to the grains in A–C. However, the higher resolution reveals that this micrometre-sized particle is itself an aggregate of smaller units; seven grains can be resolved, with sizes ranging from $260^{+50}_{-120}\text{nm}$ to $540^{+20}_{-250}\text{nm}$. The visible part of particle E has a maximum extent of $14\mu\text{m}$ in the x direction and $37\mu\text{m}$ in the y direction. Analysis of its component grains (Fig. 3d) shows sizes in the range from $0.58^{+0.15}_{-0.20}\mu\text{m}$ to $2.57^{+0.04}_{-0.51}\mu\text{m}$, with the grain heights ranging between $0.2\mu\text{m}$ and $3\mu\text{m}$ and with 90% smaller than $1.7\mu\text{m}$. These measurements are evidence for a continuation of the aggregate nature of dust particles below the size range observed by the COSIMA (Cometary Secondary Ion Mass Analyser) instrument on-board Rosetta (tens to hundreds of micrometres)¹⁵.

Particle E also shows a morphology that is strongly reminiscent of stratospheric, chondritic porous interplanetary dust particles (IDPs), which have long been suspected of having a cometary origin. This link is consistent with observations by COSIMA for larger dust particles, which also measured similar compositions for dust at comet 67P and IDPs^{15,16}. One notable difference to IDPs is the extremely flat nature of particle E, which has a height that is an order of magnitude lower than its (minimal) lateral dimension. Indeed, all of the particles presented here have flattened shapes to some degree (see Table 1). It is not yet clear if this is an intrinsic property of cometary dust or the result of a rearrangement of grains on impact. COSIMA has observed that sub-millimetre aggregate particles undergo rearrangement of their grains on impact, producing flattened shapes¹⁵. Additionally, COSIMA collected small, apparently compact particles that are also flat, but the resolution is insufficient to determine if they are single grains or aggregates. On the other hand, cluster-cluster aggregation with rotating grains can form elongated structures with very high aspect ratios¹⁷, and laboratory experiments have produced dust “flakes”¹⁸.

Investigation of the size distribution of chondritic porous IDPs and fine-grained material returned by the Stardust mission^{19,20} showed that the majority of their component grains are smaller than 500nm (refs 20, 21). Figure 3d shows that 90% of the grains in particle E are smaller than $2\mu\text{m}$, comparable to the size of particle D, which is itself

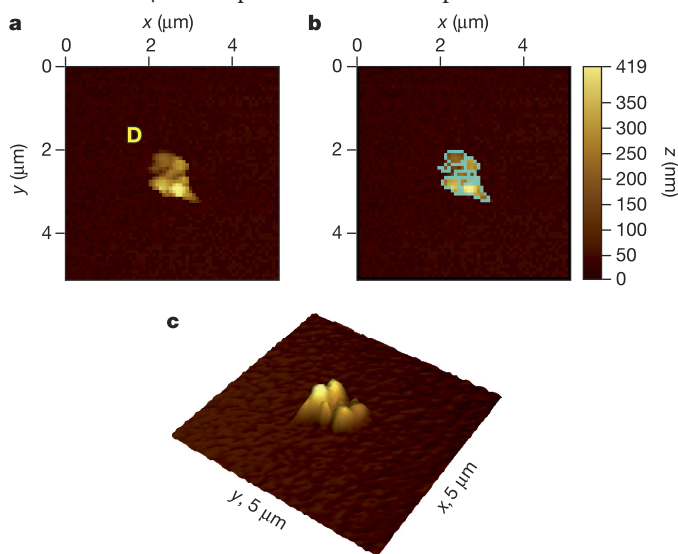


Figure 2 | AFM topographic images of particle D and its sub-units.

a, A $5\mu\text{m} \times 5\mu\text{m}$ overview image with a pixel resolution of 80nm and the colour scale representing the height, z . **b**, As in **a**, but with the sub-units of particle D outlined in cyan. **c**, A three-dimensional (rotated) image of the particle with two-times height exaggeration to aid visualization.

built from grains smaller than about 500nm . This result suggests that the grains of the fluffy aggregate particle E are also aggregates of sub-micrometre components similar to those in chondritic porous IDPs, and points towards a hierarchical structure. Hierarchical growth (that is, aggregates of smaller aggregates) has been proposed as a growth mechanism in the protoplanetary disk when fragmentation of larger particles provides a population of smaller aggregates available for agglomeration¹⁰. The sticking probability of such particles can be higher than that of homogeneous dust for a given mass and velocity and need to be accounted for in models of dust particle growth¹¹. Hierarchical aggregates have also been invoked to produce a surface layer of cometary dust with sufficiently low tensile strength to allow for dust release¹².

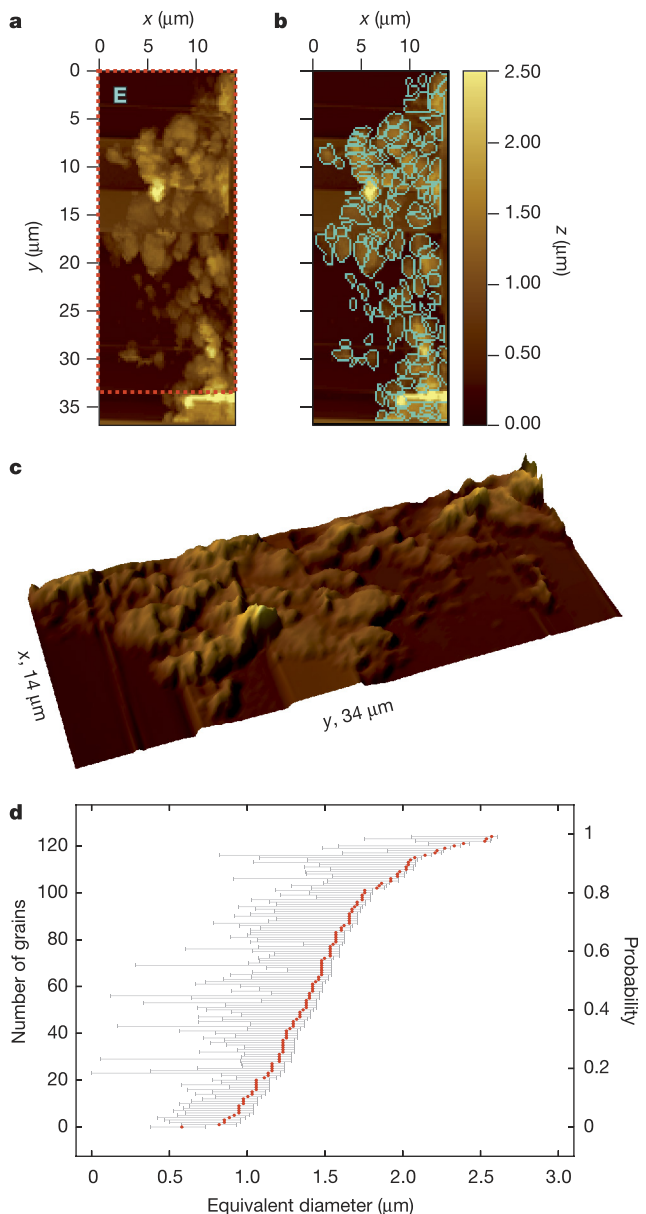


Figure 3 | AFM topographic images of particle E, showing its sub-units and their size distribution. **a**, A $14\mu\text{m} \times 37\mu\text{m}$ overview image with a pixel resolution of 210nm and the colour scale representing the height, z . **b**, As in **a**, but with identified grains outlined in cyan. **c**, A three-dimensional $14\mu\text{m} \times 34\mu\text{m}$ view (corresponding to region indicated by the red dashed box in **a**; rotated and cropped). **d**, Cumulative distribution of the equivalent diameters of the grains (red circles), with error bars in grey (where the errors are given as the linear addition of the 1σ statistical uncertainty and the systematic uncertainty; see Methods). The left scale shows the cumulative number of grains and the right scale shows the probability that particles have equivalent diameters below the specific values.

Table 1 | Size, height and elongation of dust particles A–D and their component dust grains

	Type	$d \pm \Delta d$ (μm)	z_{max} (μm)	Elongation
Particle A	Compact particle	$5.48^{+0.04}_{-1.10}$	1.79	$3.32^{+0.14}_{-0.41}$
Grain 1	Dust grain	$3.31^{+0.06}_{-1.23}$	1.79	$2.94^{+0.12}_{-0.43}$
Grain 2	Dust grain	$2.62^{+0.08}_{-0.87}$	1.33	$3.04^{+0.15}_{-0.42}$
Grain 3	Dust grain	$1.93^{+0.10}_{-1.22}$	1.57	
Grain 4	Dust grain	$2.62^{+0.08}_{-1.07}$	1.55	$1.96^{+0.09}_{-0.40}$
Particle B	Dust grain	$2.76^{+0.07}_{-0.61}$	1.02	$3.14^{+0.18}_{-0.42}$
Particle C	Compact particle	$5.79^{+0.04}_{-0.87}$	1.39	$4.77^{+0.24}_{-0.50}$
Grain 1	Dust grain	$2.66^{+0.07}_{-0.92}$	1.33	$2.26^{+0.11}_{-0.42}$
Grain 2	Dust grain	$2.57^{+0.08}_{-0.72}$	1.14	$2.80^{+0.15}_{-0.41}$
Grain 3	Dust grain	$2.18^{+0.09}_{-0.83}$	1.28	$2.23^{+0.12}_{-0.39}$
Grain 4	Dust grain	$2.42^{+0.08}_{-0.90}$	1.39	$2.31^{+0.11}_{-0.39}$
Grain 5	Dust grain	$2.31^{+0.08}_{-0.87}$	1.38	$2.32^{+0.11}_{-0.39}$
Particle D	Compact particle	$1.09^{+0.01}_{-0.25}$	0.42	$3.36^{+0.23}_{-0.47}$
Grain 1	Dust grain	$0.26^{+0.05}_{-0.12}$	0.17	$1.89^{+0.19}_{-0.36}$
Grain 2	Dust grain	$0.48^{+0.03}_{-0.16}$	0.22	$2.52^{+0.20}_{-0.47}$
Grain 3	Dust grain	$0.41^{+0.03}_{-0.14}$	0.31	$1.62^{+0.11}_{-0.27}$
Grain 4	Dust grain	$0.33^{+0.04}_{-0.13}$	0.25	$1.74^{+0.51}_{-0.71}$
Grain 5	Dust grain	$0.46^{+0.03}_{-0.17}$	0.37	$1.53^{+0.09}_{-0.28}$
Grain 6	Dust grain	$0.54^{+0.02}_{-0.25}$	0.42	$2.00^{+0.07}_{-0.82}$
Grain 7	Dust grain	$0.26^{+0.05}_{-0.15}$	0.32	$2.00^{+0.03}_{-0.97}$

d is the diameter of a circle with equivalent area to that of the particle or grain; z_{max} is the maximum height above the substrate surface. The errors of the diameters are given as the linear addition of the 1σ statistical uncertainty and the systematic uncertainty, and the errors of the elongation are given as the worst-case estimate; see Methods for further details. For particle A grain 3 and particle D grains 4, 6 and 7, the maximal elongation is found for the ratio of the two lateral dimensions that are attached with especially large uncertainties. Therefore, an accurate elongation cannot be given for particle A grain 3 and the elongations of the grains of particle D have large uncertainties.

Because MIDAS provides real measurements of the grain shapes, it is possible to evaluate which models support the observations. The elongation of the grains is found by calculating the ratio of the longest and shortest perpendicular axis. Further details are described in Methods. For particle E, the grain heights are almost all smaller than their in-plane diameters, suggesting that it comprises a single layer of grains, allowing accurate grain heights to be determined. The elongation is calculated for 114 grains (the 11 omitted grains show strong distortions due to tip convolution), giving an average elongation of $2.87^{+1.90}_{-0.44}$ (that is, the largest axis is three times longer than the smallest; the uncertainties represent a worst-case estimate containing 1σ statistical errors and systematic uncertainties). The compact particles show similar values (Table 1).

Elongated grains are considered in several models of cometary dust. For example, it has been suggested²² that comets aggregate from interstellar grains. In ref. 22, the dust grains were modelled as cylinders with aspect ratios of 2–4, and good agreement was found between light scattering experiments and observations. Other works have similarly demonstrated good agreement between simulations using aggregates of spheroidal particles and observational data^{23,24}. The elongated nature of interstellar dust can be inferred from linear polarization of starlight due to partially aligned grains²⁵. The core–mantle structure proposed for interstellar and cometary dust²⁶ cannot be confirmed by MIDAS data alone, but the elongation measurement supports the idea of a common precursor grain, or growth mechanism.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 December 2015; accepted 6 July 2016.

1. Dominik, C., Blum, J., Cuzzi, J. & Wurm, G. in *Protostars and Planets V* (eds Reipurth, B. et al.) 783–801 (Univ. Arizona Press, 2006).
2. Blum, J. Experiments on sticking, restructuring, and fragmentation of preplanetary dust aggregates. *Icarus* **143**, 138–146 (2000).
3. Dollfus, A. Polarimetry of grains in the coma of 1P/Halley. *Astron. Astrophys.* **219**, 469–478 (1989).

4. Kolokolova, L., Hanner, M. S., Levasseur-Regourd, A. C. & Gustafson, B. A. S. in *Comets II* (eds Festou, M. et al.) 577 (Univ. Arizona Press, 2004).
5. Kolokolova, L. & Kimura, H. Comet dust as a mixture of aggregates and solid particles: model consistent with ground-based and space-mission results. *Earth Planets Space* **62**, 17–21 (2010).
6. Hanner, M. S. The scattering properties of cometary dust. *J. Quant. Spectrosc. Radiat. Transf.* **79–80**, 695–705 (2003).
7. Brownlee, D. E. Morphological, chemical and mineralogical studies of cosmic dust [and discussion]. *Phil. Trans. R. Soc. A* **323**, 305–311 (1987).
8. Flynn, G. J. et al. Elemental compositions of comet 81P/Wild 2 samples collected by Stardust. *Science* **314**, 1731–1735 (2006).
9. Brownlee, D. et al. Comet 81P/Wild 2 under a microscope. *Science* **314**, 1711–1716 (2006).
10. Dominik, C. Physical processes : dust coagulation and fragmentation. in *ASP Conf. Ser. Vol. 414* (eds Henning, T. et al.) 494–508 (Astronomical Society of the Pacific, 2009).
11. Johansen, A. et al. in *Protostars and Planets VI* (eds Beuther, H. et al.) 547–570 (Univ. Arizona Press, 2014).
12. Skorov, Y. & Blum, J. Dust release and tensile strength of the non-volatile layer of cometary nuclei. *Icarus* **221**, 1–11 (2012).
13. Bentley, M. S. et al. MIDAS : lessons learned from the first spaceborne atomic force microscope. *Acta Astronaut.* **125**, 11–21 (2016).
14. Riedler, W. et al. MIDAS — the micro-imaging dust analysis system for the Rosetta mission. *Space Sci. Rev.* **128**, 869–904 (2007).
15. Langevin, Y. et al. Typology of dust particles collected by the COSIMA mass spectrometer in the inner coma of 67P/Churyumov Gerasimenko. *Icarus* **271**, 76–97 (2016).
16. Hilchenbach, M. et al. Comet 67P/Churyumov–Gerasimenko: close-up on dust particle fragments. *Astrophys. J.* **816**, L32 (2016).
17. Paszun, D. & Dominik, C. The influence of grain rotation on the structure of dust aggregates. *Icarus* **182**, 274–280 (2006).
18. Stephens, J. R. & Gustafson, B. A. S. Laboratory reflectance measurements of analogues to “dirty” ice surfaces on atmosphereless solar system bodies. *Icarus* **94**, 209–217 (1991).
19. Rotundi, A. et al. Combined micro-Raman, micro-infrared, and field emission scanning electron microscope analyses of comet 81P/Wild 2 particles collected by Stardust. *Meteorit. Planet. Sci.* **43**, 367–397 (2008).
20. Stodolna, J., Gainsforth, Z., Butterworth, A. L. & Westphal, A. J. Characterization of preserved primitive fine-grained material from the Jupiter family comet 81P/Wild 2 — a new link between comets and CP-IDPs. *Earth Planet. Sci. Lett.* **388**, 367–373 (2014).
21. Wozniakiewicz, P. J., Bradley, J. P., Ishii, H. A., Price, M. C. & Brownlee, D. E. Pre-accretionary sorting of grains in the outer solar nebula. *Astrophys. J.* **779**, 164 (2013).
22. Greenberg, J. M. & Gustafson, B. A. S. A comet fragment model for zodiacal light particles. *Astron. Astrophys.* **93**, 35–42 (1981).
23. Levasseur-Regourd, A. C., Mukai, T., Lasue, J. & Okada, Y. Physical properties of cometary and interplanetary dust. *Planet. Space Sci.* **55**, 1010–1020 (2007).
24. Lasue, J. & Levasseur-Regourd, A. C. Porous irregular aggregates of sub-micron sized grains to reproduce cometary dust light scattering observations. *J. Quant. Spectrosc. Radiat. Transf.* **100**, 220–236 (2006).
25. Mathis, J. S. Interstellar dust and extinction. *Annu. Rev. Astron. Astrophys.* **28**, 37–70 (1990).
26. Greenberg, J. M. & Hage, J. I. From interstellar dust to comets: a unification of observational constraints. *Astrophys. J.* **361**, 260–274 (1990).

Acknowledgements Rosetta is an ESA mission with contributions from its member states and NASA. We also thank the Rosetta Science Ground Segment and Mission Operations Centre for their support in acquiring the presented data. MIDAS became possible through support from funding agencies including the European Space Agency’s PRODEX programme, the Austrian Space Agency, the Austrian Academy of Sciences and the German funding agency DARA (later DLR). A.-C.L.-R. acknowledges support from the French Space Agency, CNES. M.S.B. and T.M. acknowledge funding from the Austrian Science Fund (FWF): P 28100-N36. T.M. also acknowledges the Steiermärkische Sparkasse and the Karl-Franzens Universität Graz for their financial support. P.E. acknowledges support from the NASA Astrobiology Institute. R.S. thanks F. Hofer and H. Plank for discussions and the Austrian Research Promotion Agency (FFG) for financial support. All data presented here will be made available in the ESA Planetary Science Archive (<http://www.cosmos.esa.int/web/psa/rosetta>).

Author Contributions R.S., T.M. and M.S.B. planned the experiments on MIDAS, analysed and interpreted the data and wrote the manuscript. M.S.B. developed the planning and data processing pipelines. R.S. and T.M. implemented the elongation calculations. R.S. performed the post-processing and calibration as well as the particle/grain measurement and is responsible for the graphical data presentation. T.M. considered the uncertainties for all data. A.-C.L.-R. provided information on cometary dust derived from polarimetric observations and its interpretation. H.J. supported the experiments with software updates. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.S.B. (mark.bentley@oeaw.ac.at).

Reviewer Information Nature thanks M. Fulle and L. Kolokolova for their contribution to the peer review of this work.

METHODS

Data acquisition and calibration. Exposure durations and times were planned by estimating the dust flux using the predicted spacecraft position, pointing and a dust flux model for comet 67P derived from observational data²⁷. For a graphical visualization of the exposure geometries, see Extended Data Figs 1–3.

MIDAS operates in a slightly different way from most terrestrial AFMs, by making a careful approach to the sample at each pixel position and then moving away by a so-called retraction distance before moving to the next pixel, resulting in long scan times and possible distortion^{13,14}. Distortion correction is performed using scans of on-board calibration targets, and polynomial background correction is used to remove height drifts. This procedure was performed with the data used to produce Figs 1 and 3. The scan shown in Fig. 2 was much shorter and no substantial distortion was observed; hence, only background subtraction was performed. Particle and grain heights are measured relative to the substrate surface, which is very clear for Figs 1 and 2, but the zero reference level had to be set manually for each grain in Fig. 3, because the steps would otherwise distort the measurements.

The lateral extent of particles and grains is characterized by an effective size (d), which is the diameter of a circle with the same area as the projection of all pixels forming the unit; unless stated otherwise, all references to size refer to this effective value. The peak height (z_{\max}) is the maximum elevation above the target for a given grain. Identification of particles and their sub-units is performed by visual inspection of the calibrated data and, when necessary, cross-sections through the three-dimensional data are used; see Extended Data Fig. 4.

For particle E (Fig. 3), a manual levelling of the surface was necessary, owing to the visible steps (imaging artefacts). Repeating this manual levelling process several times showed that the induced error was negligible. In addition, the height of a grain can be measured precisely only if the grain is directly on the surface and not on another grain. For particle E, most of the grains seem to fulfil this requirement, because the mean heights of the grains are smaller than their mean diameters.

Error analysis. In principle, because AFM tips cannot be infinitely sharp, the size of every particle is overestimated, owing to the tip sample convolution (that is, the

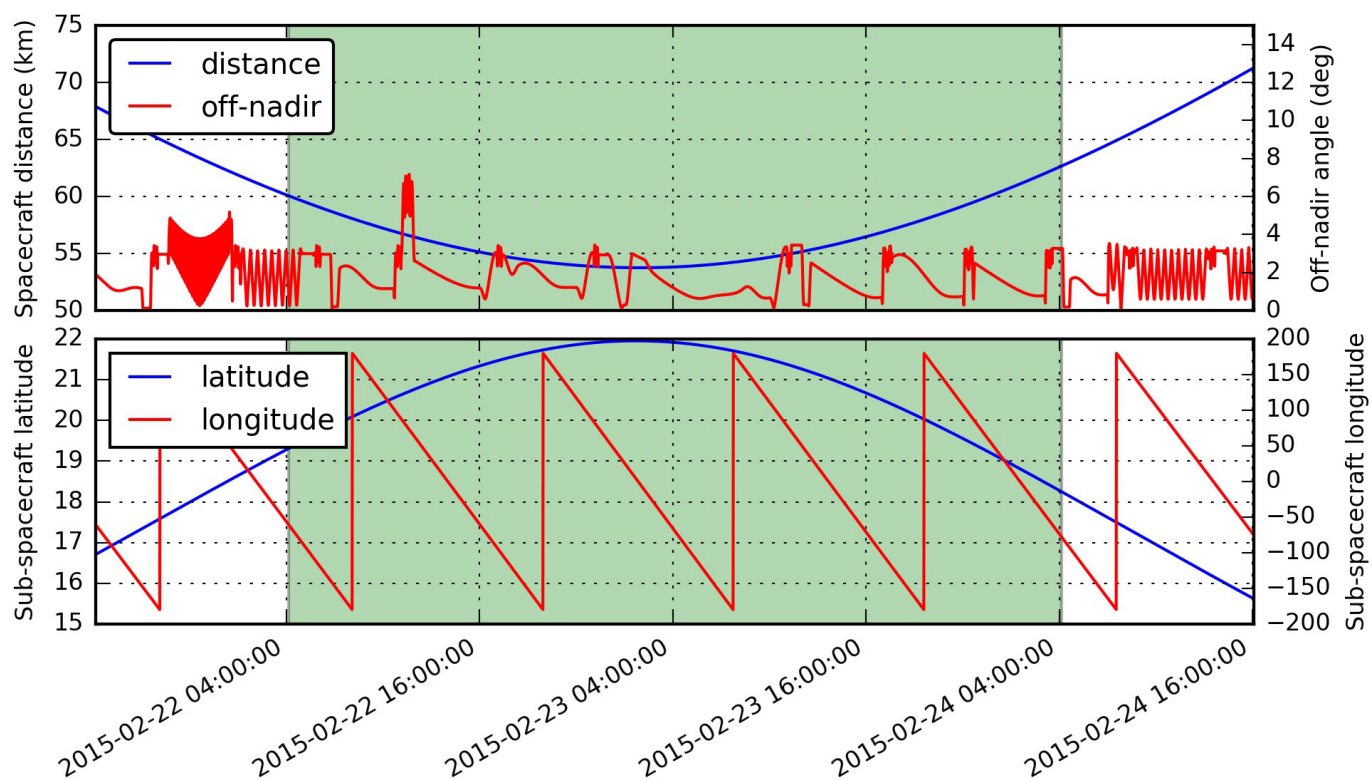
recorded image reflects a combination of tip and sample shapes). For example, a cone-shaped tip with a large opening angle artificially broadens features, as depicted in Extended Data Fig. 5. The convolution uncertainty is generously estimated here to give an upper limit. Because the particle diameter cannot be underestimated by this convolution, the uncertainty interval becomes asymmetric. This systematic uncertainty is linearly added to the 1σ statistical uncertainty generated by the identification of the grains in the scan. Values for sizes and respective uncertainties quoted in the text for all particles, depicted in Fig. 3d for particle E and presented in Table 1 for particles A–D, reflect this calculation.

The elongation of particles and grains is calculated by determining their equivalent ellipse (the ellipse with the same second-order moments) and choosing the maximum ratio of the largest to smallest of (i) the height of the particle to the major axis, (ii) the height to the minor axis and (iii) the ratio of the major and minor axes. The uncertainties in these ratios take into account the 1σ statistical uncertainty due to the manual masking of the particles and the systematic uncertainty due to the tip–sample convolution for the axis lengths. The ratio of the major to minor axis suffers from a large convolution uncertainty that, in some cases (typically particles with steep slopes), prevents a clear statement about the orientation. In these cases, no elongation is given. The final uncertainty for the ratio is a worst-case estimate that overestimates the uncertainty for non-isolated flat grains.

Code and data availability. Extended Data Table 1 summarizes the key parameters for the AFM scans used to produce Figs 1–3. The filenames listed refer to products available in the ESA Planetary Science Archive where all data used here are freely available (<http://www.cosmos.esa.int/web/psa/rosetta>). The open-source package Gwyddion²⁸ was used to perform calibration, grain identification and analysis throughout this paper (<http://gwyddion.net/>).

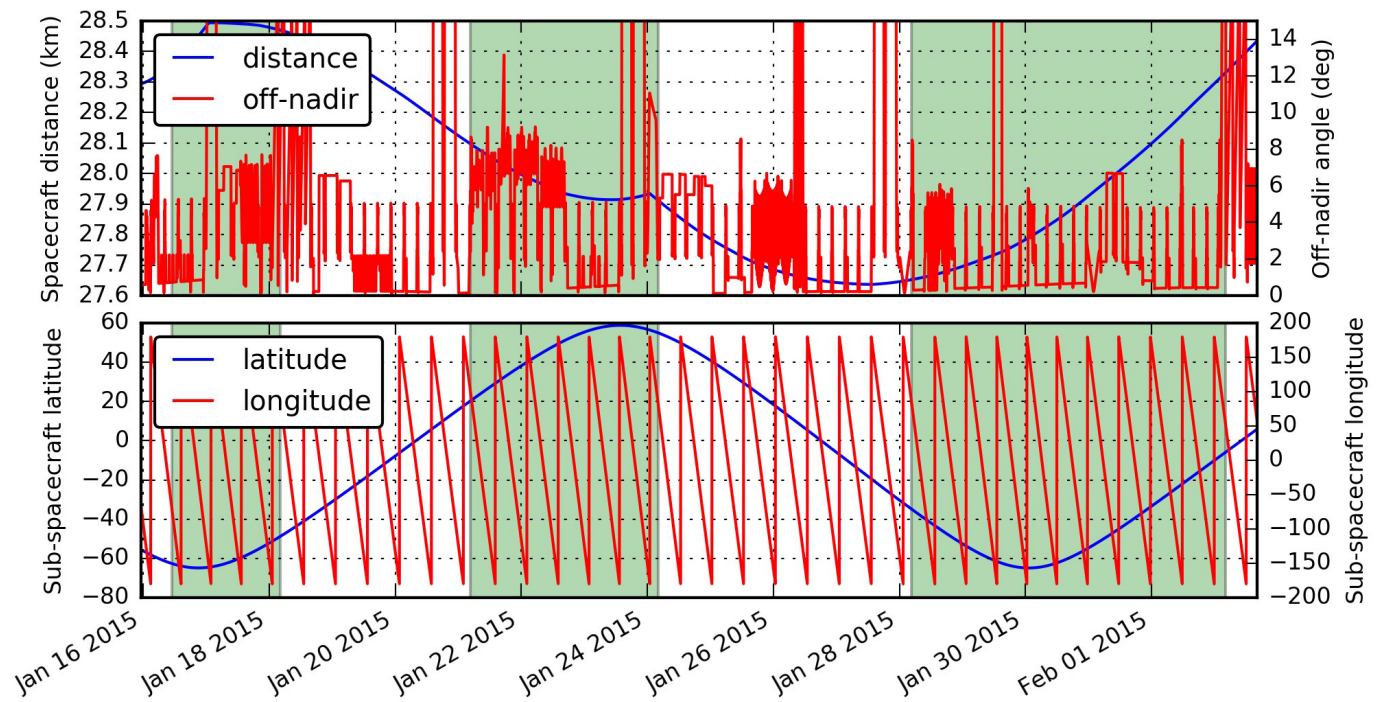
27. Fulle, M. *et al.* Comet 67P/Churyumov–Gerasimenko: the GIADA dust environment model of the Rosetta mission target. *Astron. Astrophys.* **522**, A63 (2010).

28. Nečas, D. & Klapetek, P. Gwyddion: an open-source software for SPM data analysis. *Cent. Eur. J. Phys.* **10**, 181–188 (2012).

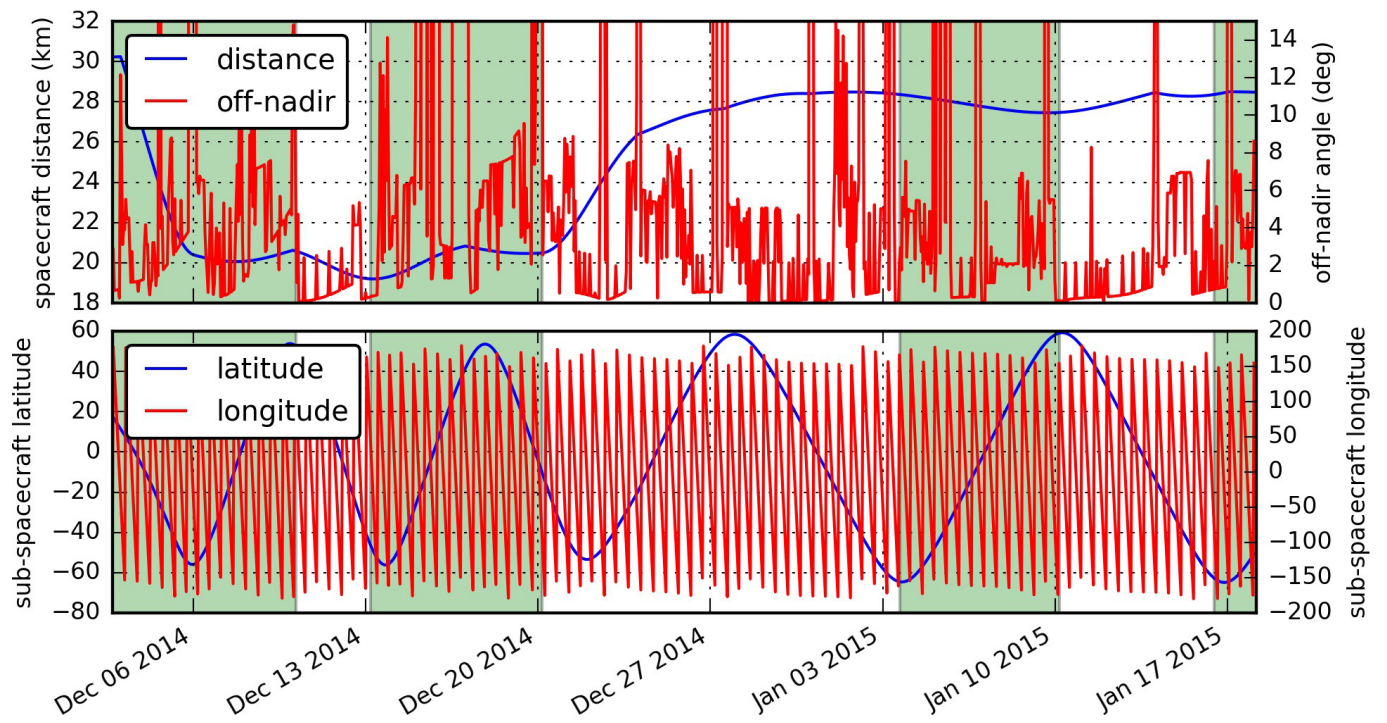


Extended Data Figure 1 | The geometry of the exposures where particles A, B and C were collected. All exposures are marked by green bars. The top panel shows the distance of Rosetta from the comet (blue) and the off-nadir angle (red). The lower panel shows the latitude

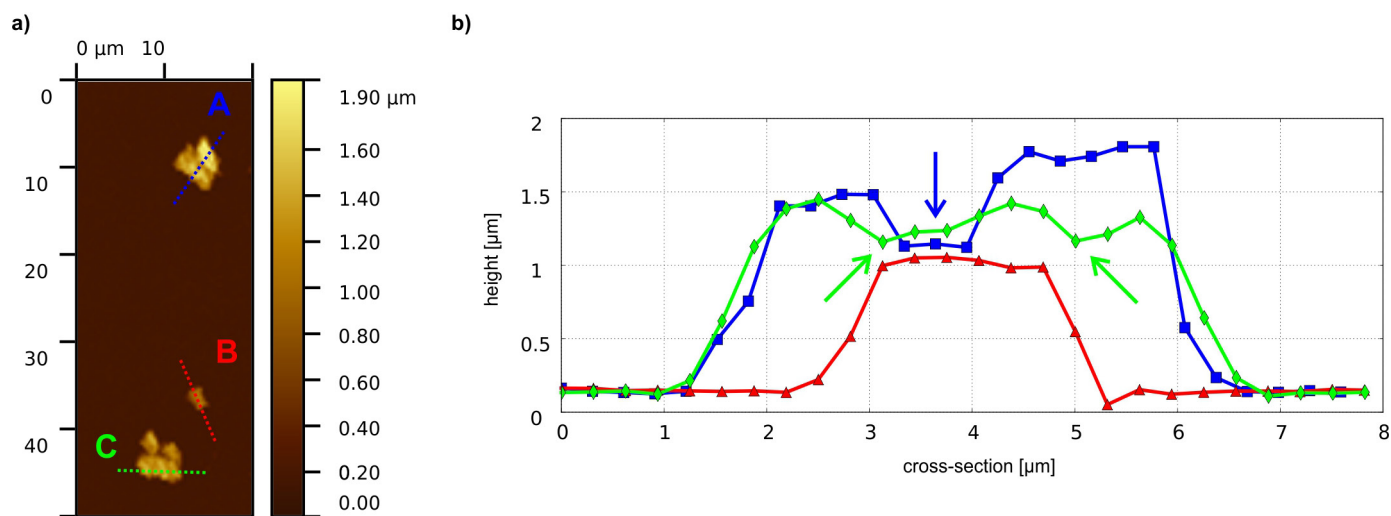
and longitude of the point on the comet below the spacecraft (the sub-spacecraft latitude and longitude) in blue and red, respectively. The heliocentric distance (between the comet and the Sun) during this exposure was 2.25 AU.



Extended Data Figure 2 | The geometry of the exposures where particle D was collected. All exposures are marked by green bars. The top panel shows the distance of Rosetta from the comet (blue) and the off-nadir angle (red). The lower panel shows the sub-spacecraft latitude and longitude in blue and red, respectively. The heliocentric distance during this exposure varied between 2.54 AU and 2.41 AU.

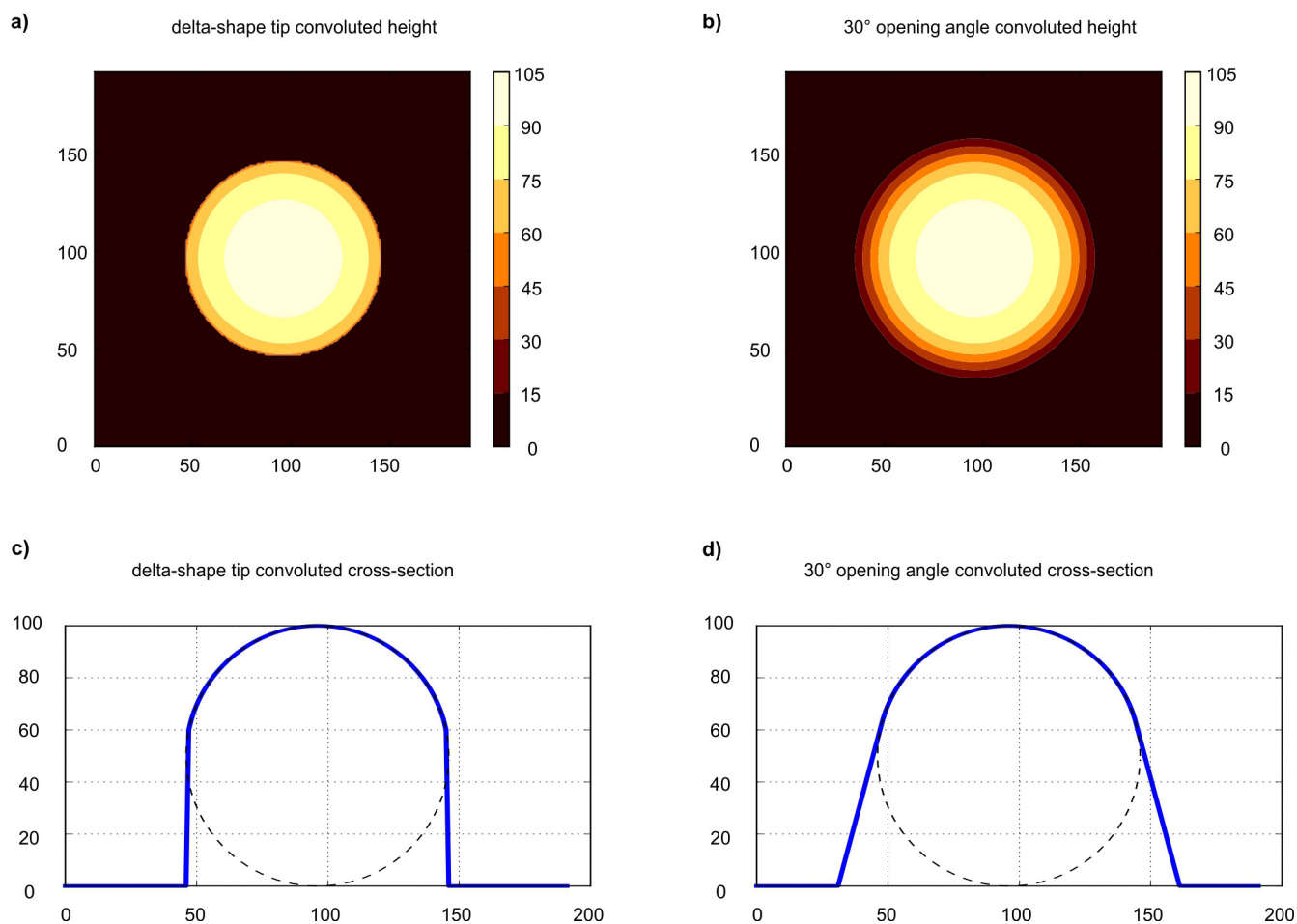


Extended Data Figure 3 | The geometry of the exposures where particle E was collected. All exposures are marked by green bars. The top panel shows the distance of Rosetta from the comet (blue) and the off-nadir angle (red). The lower panel shows the sub-spacecraft latitude and longitude in blue and red, respectively. The heliocentric distance during this exposure varied between 2.85 AU and 2.52 AU.



Extended Data Figure 4 | Topographic cross-sections demonstrating the identification of sub-units. **a**, Topographic image of particles A, B and C. Dashed blue, red and green lines show where the cross-sections of particle A, B and C, respectively, were made. The colour scale represents

the height. **b**, Height profiles of the three cross-sections shown in **a**, demonstrating how sub-grains were identified (blue and green arrows) and revealing slopes of 60° – 70° with the substrate surface.



Extended Data Figure 5 | Tip-sample convolution effects.

a, b, Simulated AFM images (colour scale indicates the height) providing a comparison between a spherical particle imaged with an ideal, infinitely sharp tip (**a**) and with a cone-shaped tip with an opening angle of 30° (**b**), which is similar to that of the MIDAS tips¹⁴. **c, d,** The corresponding cross-sections through the centre of the structures (*y* axis shows the height).

The black dashed curves show the spherical particle and the blue lines depict the topography as measured with infinitely sharp and cone-shaped tips, respectively. The measurement of the volume of the spherical particle is exaggerated by 25% for the delta-shaped tip and by 50% for the cone-shaped tip. The maximum height measurement is not affected by the tip-sample convolution.

Extended Data Table 1 | Scan parameters of the primary AFM topography scans shown in Figs 1–3

	Figure 1	Figure 2	Figure 3
target	14	12	12
cantilever	9	9	7
image resolution	256 x 256	256 x 256	192 x 192
image size	80 x 80 μm^2	20 x 20 μm^2	40 x 40 μm^2
pixel resolution	312 nm	80 nm	210 nm
z step size	0.7 nm	0.7 nm	0.7 nm
retraction height	1095 nm	977 nm	734 nm
duration	1 day, 05:05:33	08:14:15	11:16:30
start time	2015-04-29T05:21:40Z	2015-03-13T08:44:38Z	2015-01-18T20:59:28Z
filename	IMG_1509813_1512600_054_ZS	IMG_1507001_1508813_005_ZS	IMG_1501323_1504200_013_ZS

The number of pixels and the pixel resolution at a given scan size was limited by the time available and chosen to maximize the resolution. The filename corresponds to that used in the Planetary Science Archive.

Dynamically encircling an exceptional point for asymmetric mode switching

Jörg Doppler¹, Alexei A. Mailybaev², Julian Böhm³, Ulrich Kuhl³, Adrian Girschik¹, Florian Libisch¹, Thomas J. Milburn⁴, Peter Rabl⁴, Nimrod Moiseyev⁵ & Stefan Rotter¹

Physical systems with loss or gain have resonant modes that decay or grow exponentially with time. Whenever two such modes coalesce both in their resonant frequency and their rate of decay or growth, an ‘exceptional point’ occurs, giving rise to fascinating phenomena that defy our physical intuition^{1–6}. Particularly intriguing behaviour is predicted to appear when an exceptional point is encircled sufficiently slowly^{7,8}, such as a state-flip or the accumulation of a geometric phase^{9,10}. The topological structure of exceptional points has been experimentally explored^{11–13}, but a full dynamical encircling of such a point and the associated breakdown of adiabaticity^{14–21} have remained out of reach of measurement. Here we demonstrate that a dynamical encircling of an exceptional point is analogous to the scattering through a two-mode waveguide with suitably designed boundaries and losses. We present experimental results from a corresponding waveguide structure that steers incoming waves around an exceptional point during the transmission process. In this way, mode transitions are induced that transform this device into a robust and asymmetric switch between different waveguide modes. This work will enable the exploration of exceptional point physics in system control and state transfer schemes at the crossroads between fundamental research and practical applications.

Exceptional points (EPs), also called non-Hermitian degeneracies or branch points, have turned out to be at the origin of many counter-intuitive phenomena appearing in physical systems that experience gain or loss^{1–6}. Such external influences on a system require a non-Hermitian description that incorporates non-conservation of energy resulting from an external input or output. Rather than being merely a perturbative correction, gain and loss can entirely turn the behaviour of a system upside down when approaching an EP. Consider here, for example, the recent demonstrations of unidirectional invisibility^{22–24}, loss-induced suppression and revival of lasing^{25–27}, and single-mode lasers with gain and loss^{28,29} or directional output³⁰, all of which were realized at or close to an EP. These studies already nicely demonstrate the potential of EPs for novel effects and devices, but the full capability of EPs can be accessed when the EP is not just approached or swept across, but dynamically encircled^{7,8}.

Originally, it was believed that a slow encircling of an EP would result in an adiabatic evolution of states and a corresponding state flip⁹, but more recent work has rigorously shown that the same non-Hermitian components necessary for the observation of an EP actually prevent an application of the adiabatic theorem^{14–21}. Instead, non-adiabatic transitions lead to a chiral behaviour, in the sense that encircling an EP in a clockwise or a counter-clockwise direction results in different final states^{14,18,21}. While this fascinating feature has great potential for quantum control and switching protocols, it has so far defied any experimental realization. This is because to observe the non-adiabatic contributions requires a fully dynamical encircling of the EP that goes beyond the quasi-static experiments reported so far^{11–13}. A dynamically

resolved experiment is, however, extremely challenging, because of the precise control required of the two exponentially amplified or damped resonant modes that meet at the EP, which must also be decoupled from all other modes present in a system.

Proposals to overcome this problem have meanwhile been put forward, such as to map the dynamical encircling of an EP to the polarization evolution in a stratified non-transparent medium¹⁵, but the implementation requirements involved prevented an experimental realization for this case too. Here, we overcome such difficulties by demonstrating that waveguides with two transverse modes can be suitably engineered such that the transmission through them is equivalent to a slow dynamical encircling of an EP. In this way we make the recently discussed dynamical features of EPs directly accessible through established waveguide technology as used for the transmission of sound, light, microwaves and matter waves.

An EP arises when an open system described by the Schrödinger-type equation $i\partial_t\psi = H\psi$ features two resonant modes that coalesce. Such a scenario can conveniently be captured by the following non-Hermitian 2×2 Hamiltonian:

$$H = \begin{pmatrix} \delta - i\gamma_1/2 & g \\ g & -i\gamma_2/2 \end{pmatrix} \quad (1)$$

where g denotes the coupling and δ the detuning; γ_1 and γ_2 are the respective loss rates of the two relevant modes. At the specific parameter configuration $\delta_{\text{EP}} = 0$ and $g_{\text{EP}} = |\gamma_1 - \gamma_2|/4$, both the eigenvalues and eigenvectors of this Hamiltonian coalesce, which is the hallmark of the EP. As shown in Fig. 1, the vicinity of this point exhibits a characteristic structure of a self-intersecting Riemann surface. The EP marks the branch point (at the centre of each panel in Fig. 1a, b) at which the Riemann surface splits. It is this topological structure that allows one to encircle the EP such that the two eigenmodes interchange: for such a state-flip two system parameters need to be continuously changed in time t (for example, the coupling $g = g(t)$ and the detuning $\delta = \delta(t)$) along a closed loop in parameter space around the EP. This system evolution is described by the now time-dependent Hamiltonian (1) in the corresponding Schrödinger-type equation $i\partial_t\psi(t) = H(t)\psi(t)$. If the system dynamics is fully adiabatic, a flip between the two states is realized upon encircling the EP such that the lower state becomes the upper one (Fig. 1a, left). As was found only recently¹⁴, however, contributions due to the breakdown of adiabaticity in non-Hermitian systems always enter dominantly whenever both encircling directions are considered. In the case above, traversing the same parameter loop in the opposite direction thus leads to the situation that the lower state returns to itself rather than to the upper state (Fig. 1b, left). This enforces an overall asymmetric behaviour such that the state that is selected at the end of a loop depends only on the loop’s encircling direction, but not on its starting point—compare Fig. 1a and Fig. 1b for a

¹Institute for Theoretical Physics, Vienna University of Technology (TU Wien), Vienna, A-1040, Austria. ²Instituto Nacional de Matemática Pura e Aplicada—IMPA, 22460-320 Rio de Janeiro, Brazil.

³Laboratoire de Physique de la Matière Condensée, CNRS UMR 7336, Université Nice Sophia Antipolis, 06108 Nice, France. ⁴Vienna Center for Quantum Science and Technology, Atomintstitut, Vienna University of Technology (TU Wien), Vienna A-1020, Austria. ⁵Schulich Faculty of Chemistry and Faculty of Physics, Technion—Israel Institute of Technology, Haifa, 32000, Israel.

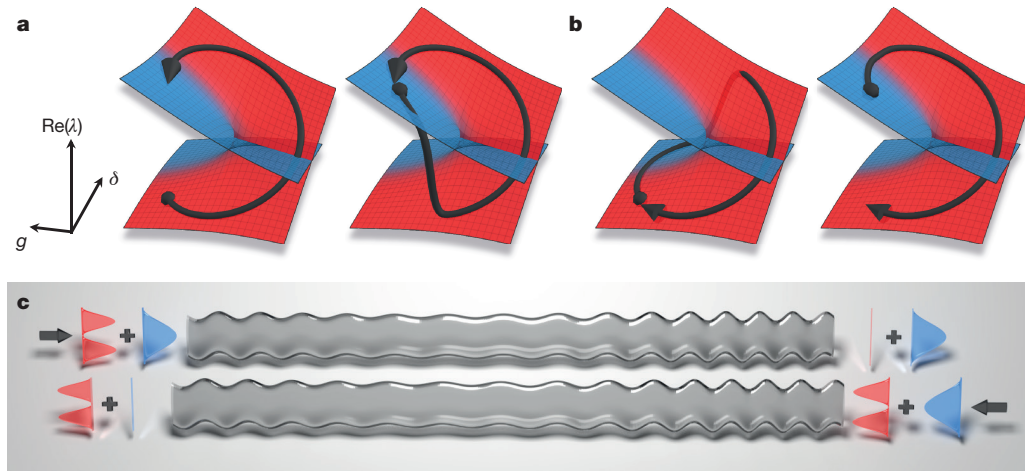


Figure 1 | Mode evolution in the vicinity of an exceptional point. To demonstrate the non-adiabatic nature of dynamically encircling an EP degeneracy, we show trajectories with different encircling directions, starting on both of the Riemann sheets involved (shown as red and blue surfaces). The results for the state evolution of the Schrödinger-type equation $i\partial_x\psi(x) = H(x)\psi(x)$, projected onto their respective Riemann sheets, are shown as black lines: the larger the contribution of an eigenvector, the closer it follows the corresponding eigensheet²¹. **a**, Dynamics of two states with starting points on different sheets

counter-clockwise and clockwise encircling, respectively. On a very fundamental level, these features are connected with the Stokes phenomenon of asymptotics^{15,16} as well as with the theory of singular perturbations and stability loss delay²¹—concepts that can be used to determine whether and when the non-adiabatic excursions shown in Fig. 1a, b occur, as well as to estimate their latest possible onset time²¹.

To observe this behaviour in a realistic environment, we now map the Hamiltonian in equation (1) onto the problem of microwave transmission through a smoothly deformed metallic waveguide in the presence of absorption (see Fig. 1c). The waveguide is extended along the x axis and we restrict the following discussion to a single transverse dimension y . Within this framework, the parametrical encircling of the EP from the 2×2 model shown above translates to a slow variation of a periodic boundary modulation along the waveguide. Directly at the EP, both the Bloch wavenumbers K and the Bloch modes A of the electric field distribution $\phi(x, y) = A(x, y)e^{iKx}$ coalesce. More specifically, the harmonic solutions $\varphi(x, y, t) = \phi(x, y)e^{-i\omega t}$ for fields oscillating with frequency ω obey the Helmholtz equation:

$$\Delta\phi(x, y) + V(x, y)\phi(x, y) = 0 \quad (2)$$

where Δ is the Laplace operator in two dimensions, $V(x, y) = \varepsilon(x, y)\omega^2/c^2$ is a complex potential proportional to the dielectric function ε , and c is the speed of light. For a straight rectangular waveguide with a fixed width W in the y direction the solutions of equation (2) in the absence of losses are $\phi_n(x, y) = u_n(y)e^{ik_n x}$ with transverse mode functions $u_n(y) = \sin(n\pi y/W)$ and wavevectors $k_n = \sqrt{\omega^2/c^2 - n^2\pi^2/W^2}$. By choosing an appropriate input frequency ω , the transmission problem can naturally be reduced to only two propagating modes $n = 1, 2$. To implement a controlled coupling between these modes, we consider a waveguide subject to a boundary modulation $\xi(x) = \sigma \sin(k_b x)$, as shown in Fig. 1c. By choosing the boundary wavenumber $k_b = k_1 - k_2 + \delta$, where $|\delta| \ll k_b$, near-resonant scattering between the otherwise very different modes ϕ_1 and ϕ_2 occurs. The full solution for the propagating field can be written in the form:

$$\phi(x, y) = \alpha_1(x)\phi_1(x, y) + \alpha_2(x)\phi_2(x, y) \quad (3)$$

Employing a Floquet–Bloch ansatz, we obtain a Schrödinger-type equation for the slowly varying modal amplitudes $\psi(x) = (c_1(x), c_2(x))^T = e^{-i\delta x}(\sqrt{ik_1}\alpha_1(x), \sqrt{-ik_2}\alpha_2(x))^T$:

during a counter-clockwise loop around the EP (as seen from the top). **b**, Same as **a** for a clockwise loop. In both **a** and **b** the end points of the loops depend only on the encircling direction, not on their starting point. **c**, Schematic of an asymmetric mode-switch that projects the above EP-encircling to a waveguide that strongly attenuates one of its two transverse modes, depending on the injection direction. The parameter-space trajectories describing counter-clockwise and clockwise loops around the EP shown in **a** and **b** correspond to the left and right injection, respectively.

$$i\partial_x \begin{pmatrix} c_1(x) \\ c_2(x) \end{pmatrix} = \begin{pmatrix} \delta(x) - i\gamma_1/2 & g(x) \\ g(x) & -i\gamma_2/2 \end{pmatrix} \begin{pmatrix} c_1(x) \\ c_2(x) \end{pmatrix} \quad (4)$$

(See Supplementary Information for a more detailed derivation verified by numerical simulations.) The slow variation of $\delta = \delta(x)$ and $g = g(x) \propto \sigma(x)$ in Hamiltonian (1) is then directly implemented in the waveguide through a smooth variation of the modulation potential $V(x, y)$, which leaves the validity of equations (3) and (4) intact. Finally, owing to the even and odd symmetry of $u_1(y)$ and $u_2(y)$, an absorbing material placed close to the centre of the waveguide gives rise to losses $\gamma_1 \gg \gamma_2$. With the above, all parameters in the non-Hermitian Hamiltonian H in equation (1) are determined. However, instead of governing the temporal dynamics (in time), H determines here the mode propagation in the longitudinal direction x . Correspondingly, the requirement of encircling the EP slowly (in time t) is transferred here to a slow variation of the boundary parameters along the propagation direction x (see Fig. 1c). Quite remarkably, a right and left injection into the waveguide corresponds to a clockwise and counter-clockwise encircling direction of the EP, respectively, yielding a specific and different output mode depending only on the side from which the waves are injected.

First numerical results following this procedure are shown in Fig. 2, where we rely on a parametrization of the waveguide modulation envelope, $\sigma(x) = (\sigma_0/2)(1 - \cos(2\pi x/L))$ that is restricted to a finite region ($x \in [0, L]$) and perfectly connects to flat semi-infinite waveguides outside this domain. Deviating from what is shown in Fig. 1, we also choose the detuning δ to be linear in x , $\delta(x) = \delta_0(2x/L - 1) + \rho$, which, together with $\sigma(x)$ from above, still describes a loop around the EP, since the endpoints of this parameter-trajectory correspond to identical waveguide configurations (see Supplementary Information for details). By implementing these design considerations in a waveguide first with uniform (bulk) loss in the transverse direction, the desired asymmetric switching of modes is, indeed, fully realized, as follows. Either mode entering from the left (Fig. 2a, b) is scattered into the first mode at the right exit lead. In contrast, any mode injected from the right side of the waveguide yields the second mode at the left exit lead (Fig. 2c, d). On the downside, however, the large overall loss both states have to acquire in order to manifest this asymmetry considerably deteriorates the quality of this switching mechanism. Additionally, the requirement of slow

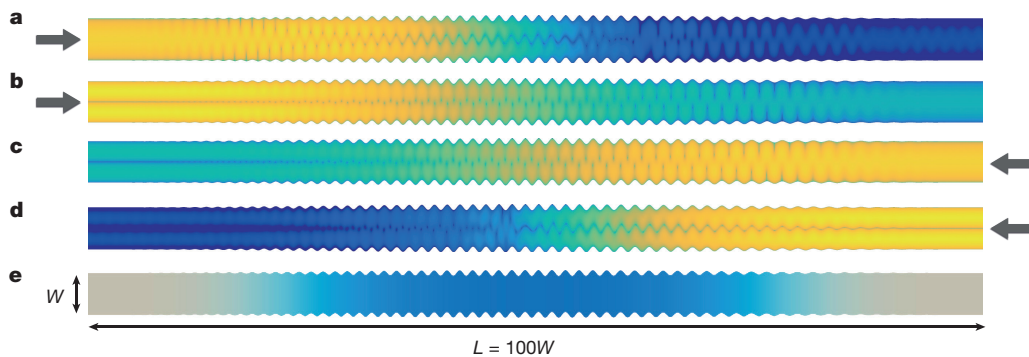


Figure 2 | Chiral transport in the presence of bulk absorption.

a–d, Numerically simulated modal wavefunction intensities for a waveguide with a length-to-width ratio of $L/W = 100$ (the depicted dimensions are not to scale). Shown are results for different input modes and injection directions: arrows indicate the side from which the waveguide is excited; the first mode is injected in **a** and **c**, and the second mode is injected in **b** and **d**. We use a logarithmic scale for the respective intensities since the overall dissipation is very strong, as is evident from

the corresponding values for transmission T_{nm} from mode n into mode m : $T_{11} = 2.5 \times 10^{-11}$, $T_{21} = 8.9 \times 10^{-7}$, $T_{12} = 7.0 \times 10^{-14}$ and $T_{22} = 8.4 \times 10^{-12}$. The normalized mode profiles at the waveguide exit, which clearly show the efficient mode-switching, are shown in Supplementary Fig. 10. **e**, Plot of the absorption strength, which is gradually switched on and off, but is uniform in the transverse direction. Specifically, for the above waveguide, $\omega W/c\pi = 2.05$, $\sigma_0/W = 0.07$, $\delta_0 W = 0.5$ and $\rho W = -0.5$.

encircling translates into a long and bulky device with many boundary oscillations.

To overcome both of these obstacles, we devised the following two strategies. First, we designed the absorption in the waveguide to follow a spatial pattern that minimizes (maximizes) the dissipation for the mode featuring the adiabatic (non-adiabatic) transition, while leaving the topology of the loop around the EP intact (see Supplementary Information for details). Remarkably, no matter which spatial profile we choose for the absorber, the reciprocity principle ensures that our design works for both transmission directions equivalently. Second, we employed a combination of quasi-Newton methods with stochastic algorithms to decrease the system length, resulting in a length-to-width ratio reduced by a factor of four as compared to the devices shown in Fig. 2. In this optimization, we tuned the parameters σ_0 , δ_0 and ρ such as to reduce the waveguide length while making sure that the resulting device still maintains the frequency robustness inherent in our design

principle (see Supplementary Fig. 11 for this efficient device geometry and the corresponding numerical results).

To demonstrate its potential for real-world applications, we provide here the first experimental realization of the above protocol, implemented in a surface-modulated microwave setup following the proposed efficient design (Fig. 3a, b). Measuring the modal transmission intensities T_{nm} from mode n into mode m as a function of the input signal frequency, we unambiguously confirm the asymmetric switching effect (see Fig. 3c): An arbitrary combination of modes injected from the left side of the waveguide is transmitted into the first mode when arriving at the exit lead on the right (T_{11} and T_{21} dominate the transmission of the first and second mode, respectively, with transmission intensity ratios $T_{11}/T_{12} = 20.6$ and $T_{21}/T_{22} = 23.0$). At the same time, the second mode is produced on the left for injection from the right (T'_{12} and T'_{22} dominate the respective modal transmission, where the primed quantities are those for

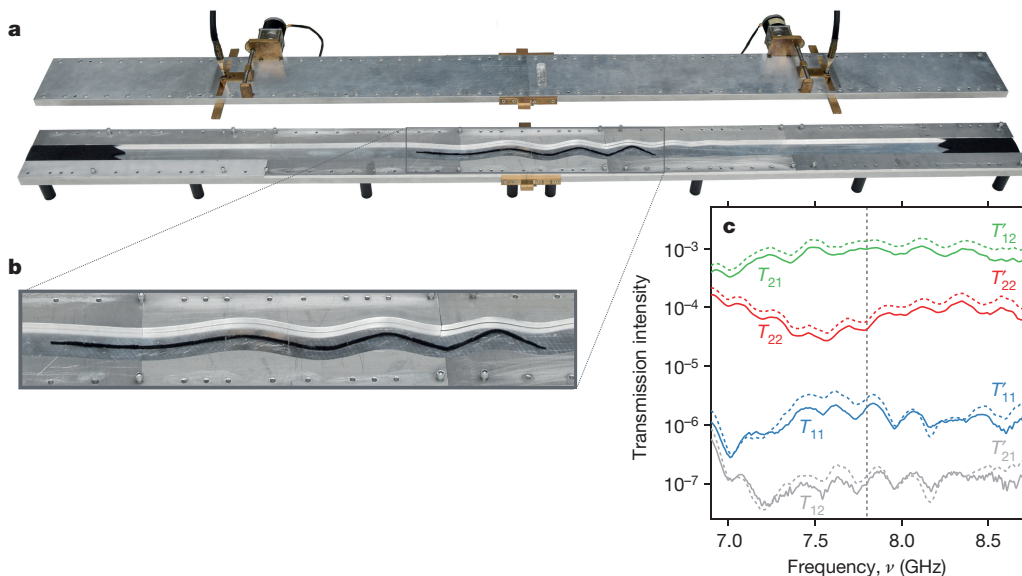


Figure 3 | Microwave measurements. **a**, Photograph of the optimized waveguide channel used in the experiment, with a surface-modulated region of length $L = 1.25$ m and width $W = 5$ cm (image credit J.B. and U.K., 2015). Within this setup, input and output antennas are placed 1.5 m apart (shown on the top plate). Black foam is used both as an absorber in the centre of the waveguide (magnified in **b**) and to mitigate the reflection into the entrance and exit leads. The setup is engineered for

a target frequency of $\nu = 7.8$ GHz (shown by a dashed vertical line in **c**), but the design ensures applicability over a broad frequency interval. **c**, Measured frequency-dependent transmission intensities T_{nm} (T'_{nm}) from mode n into mode m for injection from the left (right) are shown by solid (dashed) lines. The waveguide parameters here are $\omega W/c\pi = 2.6$, $\sigma_0/W = 0.16$, $\delta_0 W = 1.25$ and $\rho W = -1.8$.

injection from the right, with ratios $T'_{12}/T'_{11} = 463.4 \approx T_{21}/T_{11} = 488.6$ and $T'_{22}/T'_{21} = 425.9 \approx T_{22}/T_{12} = 438.4$). Note that the slight violation of the reciprocity property $T'_{nm} = T_{mn}$ observed in the experiment (see Fig. 3c) is due to the magnetized absorber material (see details in Methods), which is needed to obtain a sufficiently strong absorption in the corresponding frequency range (without the absorber, the experiment is fully reciprocal). This small non-reciprocity is, however, not essential for the operation of our device, since the respective intensity ratios are approximately the same for both injection directions. Most importantly, the experimental data proves the very strong robustness of these transmission values with respect to variations of the input frequency—a broad-band feature that is a direct consequence of our design principle, which ensures operability also in the presence of small variations of the waveguide parameters. The shortened device for which the length-to-width ratio is now $L/W = 25$ also vastly outperforms the longer device in Fig. 2 (for which $L/W = 100$), not only in terms of length-to-width ratio, but also in terms of the output intensity which is here increased by six orders of magnitude.

As an ultimate proof that the functionality of our device hinges on a dynamical EP-encircling, we also fabricated five waveguides, with individual boundary frequencies and amplitudes distributed over the parameter loop around the EP inherent in the chirped waveguide design of Fig. 3. Concatenating these stroboscopic results allows us to eliminate the dynamics in the loop around the EP, resulting in a parametric EP-encircling for which all non-adiabatic contributions should vanish. Remarkably, our results for this case (see Supplementary Fig. 9) fully reproduce the symmetric state flips that were observed in all previous experiments^{11–13} where such a parametric EP-encircling was implemented.

In summary, our work constitutes the first experimental encircling of an exceptional point that stays faithful to the full dynamical and non-adiabatic behaviour occurring in this context. In this way, we have devised a notably platform-independent approach to mode switching that is implementable not just for microwaves, but readily applicable also to light, acoustic or matter waves. An accompanying paper also reports on a dynamical EP-encircling in an optomechanical setup³¹.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 January; accepted 12 May 2016.

Published online 25 July 2016.

- Berry, M. V. Physics of nonhermitian degeneracies. *Czech. J. Phys.* **54**, 1039–1047 (2004).
- Bender, C. M. Making sense of non-Hermitian Hamiltonians. *Rep. Prog. Phys.* **70**, 947 (2007).
- Rotter, I. A non-Hermitian Hamilton operator and the physics of open quantum systems. *J. Phys. A* **42**, 153001 (2009).
- Moiseyev, N. *Non-Hermitian Quantum Mechanics* (Cambridge Univ. Press, 2011).
- Heiss, W. D. The physics of exceptional points. *J. Phys. A* **45**, 444016 (2012).
- Cao, H. & Wiersig, J. Dielectric microcavities: model systems for wave chaos and non-Hermitian physics. *Rev. Mod. Phys.* **87**, 61–111 (2015).
- Lefebvre, R., Atabek, O., Sindelka, M. & Moiseyev, N. Resonance coalescence in molecular photodissociation. *Phys. Rev. Lett.* **103**, 123003 (2009).
- Atabek, O. *et al.* Proposal for a laser control of vibrational cooling in Na₂ using resonance coalescence. *Phys. Rev. Lett.* **106**, 173002 (2011).
- Latinne, O. *et al.* Laser-induced degeneracies involving autoionizing states in complex atoms. *Phys. Rev. Lett.* **74**, 46–49 (1995).
- Mailybaev, A. A., Kirillov, O. N. & Seyranian, A. P. Geometric phase around exceptional points. *Phys. Rev. A* **72**, 014104 (2005).
- Dembowski, C. *et al.* Experimental observation of the topological structure of exceptional points. *Phys. Rev. Lett.* **86**, 787–790 (2001).

- Lee, S.-B. *et al.* Observation of an exceptional point in a chaotic optical microcavity. *Phys. Rev. Lett.* **103**, 134101 (2009).
- Gao, T. *et al.* Observation of non-Hermitian degeneracies in a chaotic exciton-polariton billiard. *Nature* **526**, 554–558 (2015).
- Uzdin, R., Mailybaev, A. & Moiseyev, N. On the observability and asymmetry of adiabatic state flips generated by exceptional points. *J. Phys. A* **44**, 435302 (2011).
- Berry, M. V. Optical polarization evolution near a non-Hermitian degeneracy. *J. Opt.* **13**, 115701 (2011).
- Berry, M. V. & Uzdin, R. Slow non-Hermitian cycling: exact solutions and the Stokes phenomenon. *J. Phys. A* **44**, 435303 (2011).
- Demange, G. & Graefe, E.-M. Signatures of three coalescing eigenfunctions. *J. Phys. A* **45**, 025303 (2012).
- Gilary, I., Mailybaev, A. A. & Moiseyev, N. Time-asymmetric quantum-state-exchange mechanism. *Phys. Rev. A* **88**, 010102 (2013).
- Graefe, E.-M., Mailybaev, A. A. & Moiseyev, N. Breakdown of adiabatic transfer of light in waveguides in the presence of absorption. *Phys. Rev. A* **88**, 033842 (2013).
- Kapralová-Žd'ánská, P. R. & Moiseyev, N. Helium in chirped laser fields as a time-asymmetric atomic switch. *J. Chem. Phys.* **141**, 014307 (2014).
- Milburn, T. J. *et al.* General description of quasiadiabatic dynamical phenomena near exceptional points. *Phys. Rev. A* **92**, 052124 (2015).
- Lin, Z. *et al.* Unidirectional invisibility induced by \mathcal{PT} -symmetric periodic structures. *Phys. Rev. Lett.* **106**, 213901 (2011).
- Regensburger, A. *et al.* Parity–time synthetic photonic lattices. *Nature* **488**, 167–171 (2012).
- Feng, L. *et al.* Experimental demonstration of a unidirectional reflectionless parity–time metamaterial at optical frequencies. *Nat. Mater.* **12**, 108–113 (2012).
- Lietzner, M. *et al.* Pump-induced exceptional points in lasers. *Phys. Rev. Lett.* **108**, 173901 (2012).
- Brandstetter, M. *et al.* Reversing the pump dependence of a laser at an exceptional point. *Nat. Commun.* **5**, 4034 (2014).
- Peng, B. *et al.* Loss-induced suppression and revival of lasing. *Science* **346**, 328–332 (2014).
- Hodaei, H., Miri, M.-A., Heinrich, M., Christodoulides, D. N. & Khajavikhan, M. Parity–time–symmetric microring lasers. *Science* **346**, 975–978 (2014).
- Feng, L., Wong, Z. J., Ma, R.-M., Wang, Y. & Zhang, X. Single-mode laser by parity–time symmetry breaking. *Science* **346**, 972–975 (2014).
- Peng, B. *et al.* Chiral modes and directional lasing at exceptional points. *Proc. Natl Acad. Sci.* **113**, 6845–6850 (2016).
- Xu, H., Mason, D., Jiang, L. & Harris, J. G. E. Topological energy transfer in an optomechanical system with exceptional points. *Nature* <http://www.dx.doi.org/10.1038/nature18604> (2016).

Supplementary Information is available in the online version of the paper.

Acknowledgements J.D., A.G. and S.R. are supported by the Austrian Science Fund (FWF) through project numbers SFB IR-ON F25-14, SFB-NextLite F49-P10 and I 1142-N27 (GePartWave). The computational results presented were achieved in part using the Vienna Scientific Cluster. A.A.M. is supported by the National Council for Scientific and Technological Development (CNPq) grant number 302351/2015-9 and by the FAPERJ grant number E-26/210.874/2014. J.B. and U.K. acknowledge ANR project number I 1142-N27 (GePartWave). F.L. acknowledges support by the FWF through SFB-F41 VI-COM. T.J.M. and P.R. are supported by the FWF through DK CoQuS W 1210, SFB FOQUS F40, START (grant number Y 591-N16), and project OPSOQI (316607) of the WWTF. N.M. acknowledges I-Core (the Israeli Excellence Center ‘Circle of Light’) and the Israel Science Foundation (grant numbers 298/11 and 1530/15) for their financial support.

Author Contributions J.D., A.A.M., A.G., F.L., T.J.M., P.R., N.M., and S.R. developed the theoretical framework and performed numerical simulations. J.B., J.D. and U.K. designed the experiment. J.B. and U.K. were responsible for the experimental implementation, the data acquisition and its evaluation. All authors contributed to the analysis, interpretation and discussion of the theoretical and experimental findings, as well as to the preparation of the manuscript. The project was jointly supervised by A.A.M. and S.R. (theory) and by U.K. (experiment).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.R. (stefan.rotter@tuwien.ac.at), A.A.M. (alexei@impa.br) (theory) and U.K. (ulrich.kuhl@unice.fr) (experiment).

METHODS

Numerical simulations. In our numerical simulations we solve the Helmholtz equation (2) on a finite-difference grid by means of a Green's function method³².

The transmission (reflection) amplitudes t_{nm} (r_{nm}) are then determined by projecting the system's Green's function onto the flux-carrying modes in the semi-infinite leads that are attached to the scattering geometry. The corresponding intensities are given by $T_{nm} = |t_{nm}|^2$ and $R_{nm} = |r_{nm}|^2$, respectively. We choose the real part of the potential $V(x, y)$ to be finite (infinite) inside (outside) the cavity, corresponding to Dirichlet boundary conditions, and the imaginary part of the potential is determined such as to satisfy the protocol described in the main text.

Experimental setup. The experimental device is an aluminium waveguide with dimensions $L \times W \times H = 2.38 \text{ m} \times 5 \text{ cm} \times 8 \text{ mm}$. Figure 3a shows the surface modulation that steers the modes around the EP. Our microwave experiment allows us to define the corresponding boundary conditions very accurately and to place the magnetized absorbing foam material (LS-10211 foam from ARC Technologies, $W \times H = 2.5 \text{ mm} \times 5 \text{ mm}$) with sub-wavelength ($< 0.5 \text{ mm}$) precision. Additional absorbers (LS-14 and LS-16 foams from Emerson and Cuming, $W \times L = 5 \text{ cm} \times 17.5 \text{ cm}$) are employed to mimic semi-infinite leads.

Microwave measurements. To probe the sinusoidal modes formed by the z component of the electric field E_z (ref. 33), we use two microwave antennas 1.5 m apart. The antennas are fixed onto motor-controlled, moveable slides and measure the complex transmission signal outside of the modulated surface area at 2×2 points along the y axis of the antennas. For the measurements we employ microwaves with a frequency around $\nu = 7.8 \text{ GHz}$, which is well below the cutoff frequency for TE_0 modes ($\nu_c = c/2H = 18.75 \text{ GHz}$), such that only the first two sinusoidal TE_0 modes contribute to the transport. By applying the twofold sine transformation:

$$t_{nm} = \frac{1}{2} \sum_{y_1, y_2} t'(y_1, y_2) \sin\left(\frac{n\pi}{W} y_1\right) \sin\left(\frac{m\pi}{W} y_2\right)$$

where $t'(y_1, y_2)$ denotes the normalized transmission measured between antenna 1 at position y_1 and antenna 2 at position y_2 , we obtain the transmission matrix t_{nm} in its mode representation. The normalization is necessary to overcome the frequency dependent coupling of the two antennas, and is given by:

$$t'(y_1, y_2) = \frac{t(y_1, y_2)}{\sqrt{(1 - |\langle r_1^a \rangle|^2)(1 - |\langle r_2^a \rangle|^2)}}$$

Here, $t(y_1, y_2)$ describes the transmission amplitude and $\langle r_1^a \rangle$ ($\langle r_2^a \rangle$) denotes the measured reflection amplitude at antenna 1 (antenna 2), averaged over all positions y_1 (y_2) and over a frequency window of 0.076 GHz. The measured reflection amplitudes are dominated by an imperfect impedance matching between the antenna and the channel. This results in a strong reflection signal originating from the antenna itself, which contains no information about the waveguide. We thus normalize the intensity fed into the system with the denominator in the above expression for $t'(y_1, y_2)$, which allows us to compare the transmission in a broader frequency range.

32. Libisch, F., Rotter, S. & Burgdörfer, J. Coherent transport through graphene nanoribbons in the presence of edge disorder. *New J. Phys.* **14**, 123006 (2012).
33. Dietz, O., Kuhl, U., Stöckmann, H.-J., Makarov, N. M. & Izrailev, F. M. Microwave realization of quasi-one-dimensional systems with correlated disorder. *Phys. Rev. B* **83**, 134203 (2011).

Topological energy transfer in an optomechanical system with exceptional points

H. Xu¹, D. Mason¹, Luyao Jiang¹ & J. G. E. Harris^{1,2}

Topological operations can achieve certain goals without requiring accurate control over local operational details; for example, they have been used to control geometric phases and have been proposed as a way of controlling the state of certain systems within their degenerate subspaces^{1–8}. More recently, it was predicted that topological operations can be used to transfer energy between normal modes, provided that the system possesses a specific type of degeneracy known as an exceptional point^{9–11}. Here we demonstrate the transfer of energy between two vibrational modes of a cryogenic optomechanical device using topological operations. We show that this transfer arises from the presence of an exceptional point in the spectrum of the device. We also show that this transfer is non-reciprocal^{12–14}. These results open up new directions in system control; they also open up the possibility of exploring other dynamical effects related to exceptional points^{15,16}, including the behaviour of thermal and quantum fluctuations in their vicinity.

An externally imposed time variation of the Hamiltonian H of an otherwise isolated, conservative system provides a powerful means for controlling the evolution of the system. If H is varied sufficiently slowly, then the adiabatic theorem states that a system prepared at some initial time t_0 in a non-degenerate normal mode of $H(t_0)$ will remain in the corresponding normal mode of the instantaneous $H(t)$ (ref. 17). As a result, varying H so as to execute a closed loop (in the space of parameters that define H) will return the system to its initial state, up to an overall phase. This phase was shown by Berry and others to include a contribution that is determined by a simple geometric property of the control loop^{1–4}. The subsequent insight that such a topological operation (that is, executing a closed control path) may have an outcome that is robust against small fluctuations in the control path has had a profound impact on many areas of theory and experiment^{5–8,18}.

More recently, it was predicted^{9–11} that topological operations may also be used to transfer energy between modes in systems that are subject to loss and/or gain. Specifically, energy transfer was predicted to occur for closed adiabatic control paths that enclose an exceptional point (EP, a form of degeneracy that can arise when the effective Hamiltonian is non-Hermitian; also known as a branch point). It was also predicted^{12–14} that such operations can be non-reciprocal in their dependence on the initial conditions of the system and the direction of rotation of the control loop about the EP. The possibility of using topological operations to control the energy distribution within a system while also inducing non-reciprocal behaviour has attracted considerable attention^{19–22}. Some features of EPs have been demonstrated in static measurements of spectra and eigenmodes^{23,24}, however, experiments have not yet realized topological or non-reciprocal dynamics by encircling an EP.

Here we measure topological and non-reciprocal dynamics in an optomechanical system. We show that the system possesses an EP and that external control parameters can be used to encircle the EP on timescales comparable to the lifetime of the excitations of the system. We demonstrate that such topological operations can transfer energy and that this energy transfer is non-reciprocal. When the control path

is not adiabatic, the dynamics becomes more complicated; however, we find quantitative agreement between experimental data and numerical simulations over the full range of measurements.

The system studied here consists of a silicon nitride membrane placed inside a high-finesse optical cavity²⁵. The dimensions of the membrane are $1\text{ mm} \times 1\text{ mm} \times 50\text{ nm}$. Because it is almost perfectly square, the vibrational eigenmodes of the membrane include nearly degenerate pairs that are well-separated in frequency from all the other eigenmodes. We use this separation to focus on a nearly degenerate pair with natural frequencies $\omega_1/(2\pi) = 788.024\text{ kHz}$ and $\omega_2/(2\pi) = 788.487\text{ kHz}$. In the absence of laser light driving the optical cavity, these two modes are essentially uncoupled and have very small damping rates ($\gamma_1/(2\pi) = 0.6\text{ Hz}$ and $\gamma_2/(2\pi) = 1.4\text{ Hz}$).

When a laser excites the cavity, the resultant intracavity field α drives the vibrations of the membrane via radiation pressure. At the same time, these vibrations detune the cavity and thereby modulate α (refs 25, 26). It is straightforward to integrate $\alpha(t)$ out of the full optomechanical equations of motion (see Methods), resulting in an effective equation of motion for just c_1 and c_2 , the displacements of the modes of the membrane:

$$i\dot{C}(t) = HC(t) \quad (1)$$

where $C(t) = [c_1(t), c_2(t)]^T$. The effective Hamiltonian is

$$H = \begin{pmatrix} \omega_1 - i\frac{\gamma_1}{2} - ig_1^2\sigma & -ig_1g_2\sigma \\ -ig_1g_2\sigma & \omega_2 - i\frac{\gamma_2}{2} - ig_2^2\sigma \end{pmatrix} \quad (2)$$

where $g_{1,2}$ are the optomechanical coupling rates of the mechanical modes, and the complex mechanical susceptibility introduced by the intracavity field is

$$\sigma = \frac{P}{\hbar\Omega_L} \frac{\kappa_{\text{in}}}{(\kappa/2)^2 + \Delta^2} \left[\frac{1}{\kappa/2 - i(\omega_0 + \Delta)} - \frac{1}{\kappa/2 + i(-\omega_0 + \Delta)} \right] \quad (3)$$

Here P and Ω_L are the power and frequency of the laser driving the cavity, Δ is the mean detuning between the laser and the cavity, $\omega_0 = (\omega_1 + \omega_2)/2$, and κ and κ_{in} are the linewidth and input coupling rate of the cavity, respectively. The experiment described here is classical; the reduced Planck constant \hbar appears in the expression for σ because $g_{1,2}$ are given in terms of the single-photon rate.

The system will possess an EP if σ can be made to equal $(\omega_1 - i\gamma_1/2 - \omega_2 + i\gamma_2/2) \left[-i(g_1^2 - g_2^2) \pm 2g_1g_2 \right] / (g_1^2 + g_2^2)^2$. Achieving this typically requires control over both $\text{Re}(\sigma)$ and $\text{Im}(\sigma)$. For optomechanical devices in the resolved sideband regime ($\kappa < \omega_0$), this control is provided by P and Δ . By contrast, when $\kappa \gg \omega_0$, P and Δ appear in σ in a linearly dependent fashion and so control only $|\sigma|$. The ability to access (and encircle) an EP using the detuning and power of a single laser is an important feature of the system presented here (and

¹Department of Physics, Yale University, New Haven, Connecticut 06511, USA. ²Department of Applied Physics, Yale University, New Haven, Connecticut 06511, USA.

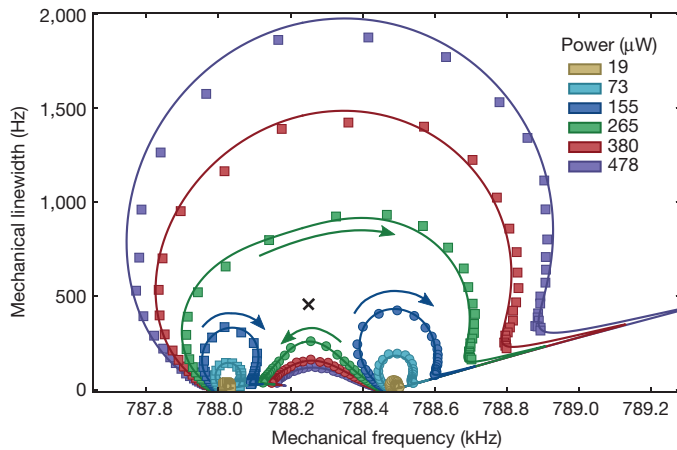


Figure 1 | The complex eigenvalues of the normal modes of the membrane. The resonance frequency (horizontal axis) and damping rate (vertical axis) of the two mechanical modes of the membrane as a function of the laser power P and detuning Δ . Data for one mode are shown as squares; data for the other mode are shown as circles. The statistical uncertainty in the measurements is smaller than the symbols. Colours indicate P , while the arrows indicate the variation of the eigenvalues as Δ is varied from $-1,200$ kHz to -400 kHz at fixed P . For the lower values of P , each eigenvalue follows a closed trajectory, beginning and ending at the same point. For the higher values of P , the eigenvalues follow open trajectories, each one ending at the starting point of the other. The solid lines are the global fit described in the text. The location of the EP predicted by this fit is shown as a black cross.

in contrast with the more complicated arrangement proposed in ref. 27), because these parameters can be controlled *in situ* with a high degree of precision, timing accuracy, and dynamic range.

A detailed description of the optomechanical device and the measurement set-up is given in Methods. The membrane and optical cavity are maintained at $T = 4.2$ K. The motion of the membrane is monitored via a heterodyne measurement of a laser with constant power and detuning. Control over the optomechanical system is

provided by a separate laser, whose detuning Δ and power P are set by an acousto-optic modulator.

To establish the presence of an EP in this system, we measured the mechanical spectrum of the membrane as a function of Δ and P . These spectra were acquired by driving the membrane and monitoring its response via the heterodyne signal. As described in Methods, each spectrum was fitted to determine the two resonance frequencies $\omega_{a,b}(\Delta, P)$ and damping rates $\gamma_{a,b}(\Delta, P)$. (The subscripts 'a' and 'b' refer to the normal modes of the membrane in the presence of an optical field; the subscripts '1' and '2' used previously refer to these modes in the absence of an optical field.)

The results of these fits are summarized in Fig. 1, which shows the complex eigenvalues $\xi_{a,b} = \omega_{a,b} - i\gamma_{a,b}/2$ as Δ and P are varied. When $P \leq 155 \mu\text{W}$, ξ_a and ξ_b each trace out a closed trajectory, completing a loop as Δ is varied from $\ll -\omega_0$ to $\gg -\omega_0$. By contrast, when $P \geq 265 \mu\text{W}$, ξ_a and ξ_b both follow open trajectories, swapping their values as Δ is varied over the same range. This sharp transition in the topology of $\xi_{a,b}(\Delta)$ is characteristic of an EP⁹. The solid lines in Fig. 1 are a global fit to the complex eigenvalues of H , which gives best-fit values of $\omega_{1,2}$ and $\gamma_{1,2}$ as stated above, as well as $g_1/(2\pi) = 1.03$ Hz, $g_2/(2\pi) = 1.14$ Hz, $\kappa_{\text{in}}/(2\pi) = 70$ kHz and $\kappa/(2\pi) = 177$ kHz. These values imply the existence of an EP at $\Delta_{\text{EP}}/(2\pi) = -792.5$ kHz, $P_{\text{EP}} = 223 \mu\text{W}$ (or equivalently $\omega_{\text{EP}}/(2\pi) = 788.2$ kHz and $\gamma_{\text{EP}}/(2\pi) = 460$ Hz, indicated as the black cross in Fig. 1).

Figure 2a, b shows measurements of $\text{Re}(\xi_{a,b})$ and $-2\text{Im}(\xi_{a,b})$ over a narrow range of Δ and P centred on Δ_{EP} and P_{EP} . These measurements show the characteristic features of an EP: ξ_a and ξ_b coalesce at a single value of the control parameters and, in the vicinity of this point, they exhibit the same structure as the Riemann sheets of the complex square-root function $z^{1/2}$. For comparison, Fig. 2c, d shows the eigenvalues of H (see equation (2)), calculated using the best-fit values determined in Fig. 1.

The surfaces shown in Fig. 2a, b are such that if Δ and P were varied to execute a single closed loop, the resulting smooth evolution on the eigenvalue manifold would return to its starting point only if the loop did not enclose the EP. By contrast, a loop enclosing the EP would result in a trajectory starting on one sheet, but ending on the other.

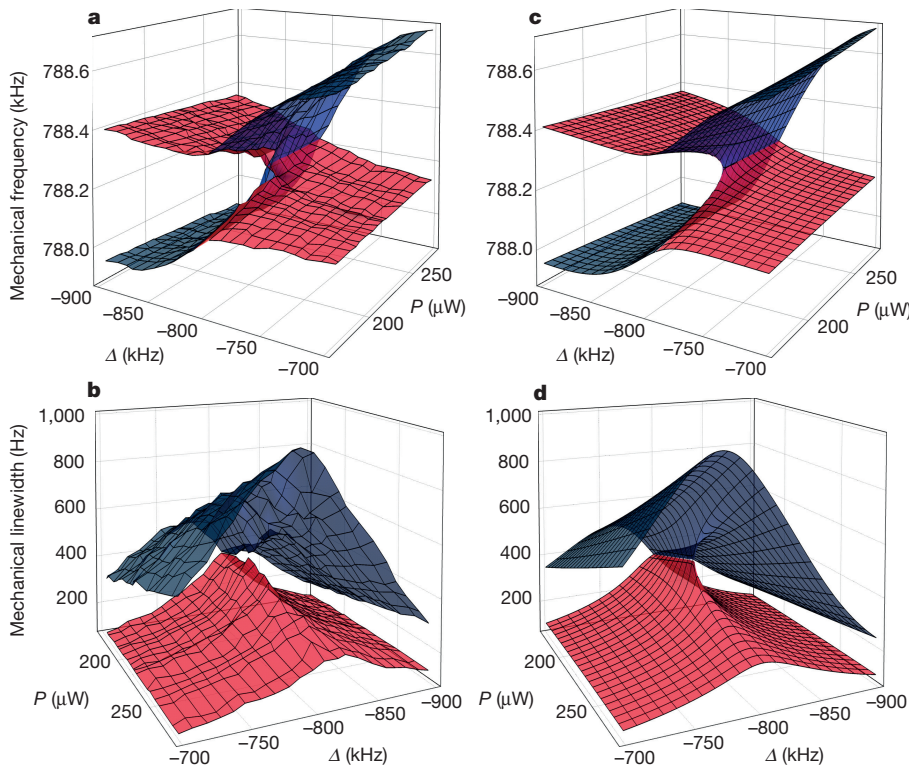


Figure 2 | The exceptional point in the spectrum of mechanical modes. a, b, The resonance frequencies (a) and damping rates (b) of the two mechanical modes of the membrane as a function of laser power P and detuning Δ . Each grid point corresponds to a measurement; grid lines and surface colouring are guides to the eye. Colouring is chosen so that red (blue) corresponds to the mode with lower (higher) damping. c, d, Plots of the theoretically calculated real (c) and imaginary (d) parts of the eigenvalues of the effective Hamiltonian matrix H (equation (2)). All of the parameters appearing in this calculation are taken from the fit in Fig. 1. Note that the viewing angle in a and c differs from that in b and d.

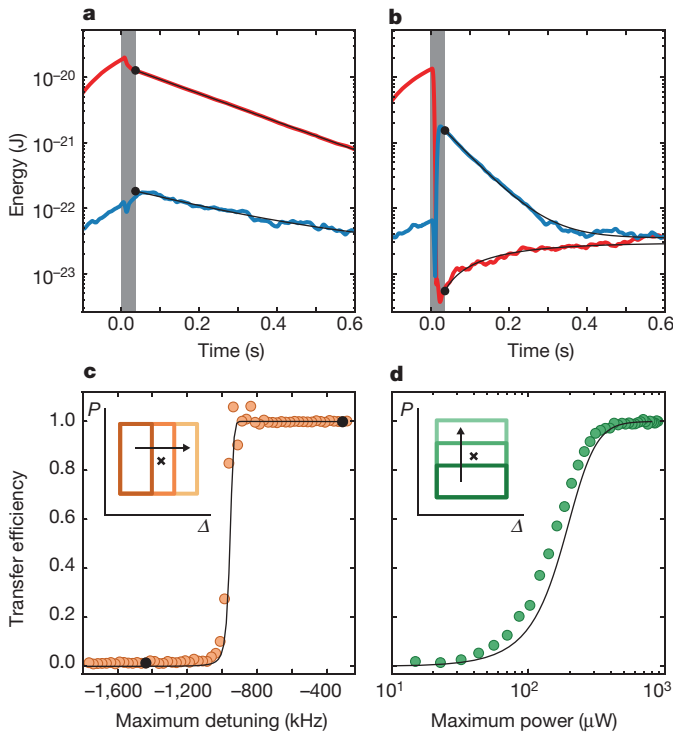


Figure 3 | Topological energy transfer. **a, b,** The energies of mode 'a' (red) and mode 'b' (blue) as a function of time t . A drive is applied to the 'a' mode for $t < 0$. At $t = 0$ the drive is turned off and the control loop described in the text is implemented. The control loop ends at $t = 16$ ms; the grey shaded region corresponds to the time during which the control loop is implemented. For $t > 16$ ms the system relaxes to thermal equilibrium. The black lines are fits to a decaying exponential (due to the mechanical damping) with a constant offset (reflecting the thermal motion of the mode). The black dot shows the extrapolation of this fit to $t = 16$ ms. The loop used in **a** does not enclose the EP, whereas the loop used in **b** does. **c,** The fraction of the (remaining) energy in the 'b' mode after the control loop has been completed as a function of the maximum detuning of the loop, Δ_{\max} . The left (right) point shown as a solid circle corresponds to the data in **a** (**b**). **d,** The corresponding measurement as a function of the maximum power of the loop, P_{\max} . In **c** and **d**, the statistical errors are comparable to or smaller than the size of the symbols. The solid lines are numerical simulations of the dynamics and are completely constrained by the parameters from the fit in Fig. 1. The insets are schematics showing how the loop varies along the horizontal axis of each panel; the location of the EP is indicated by the black cross.

To observe this effect, we performed a series of measurements in which Δ and P were initially set to Δ_{\max} and P_{\min} , and one of the modes of the membrane (c_a) was excited using a piezoelectric element. Once the system reached its steady state, the piezo drive was switched off, and Δ and P were varied to sweep out a closed rectangular loop. The loop was defined by the points $(\Delta_{\max}, P_{\min})$, $(\Delta_{\max}, P_{\max})$, $(\Delta_{\min}, P_{\max})$ and $(\Delta_{\min}, P_{\min})$, returning to $(\Delta_{\max}, P_{\min})$ after a duration $\tau = 16$ ms. This value of τ was chosen so that nearly all such control loops satisfy the requirement of conventional adiabaticity: $\tau \gg 1/|\xi_a - \xi_b|$ (loops passing close to the EP do not satisfy this inequality). We describe the effect of varying τ below.

The heterodyne signal was recorded before, during and after the control loop. This signal was demodulated at frequencies $\omega_a(\Delta_{\max}, P_{\min})$ and $\omega_b(\Delta_{\max}, P_{\min})$, with typical results shown in Fig. 3a, b. Before and after the control loop (that is, for $t < 0$ and $t > \tau$), this record corresponds to the amplitudes of the motion of the normal modes $|c_a(t)|$ (red in Fig. 3a, b) and $|c_b(t)|$ (blue). During the control loop ($0 \leq t \leq \tau$) this correspondence does not hold, because the eigenfrequencies of the membrane undergo rapid variations; data from this region do not play any role in our analysis. As shown in Fig. 3a, b, c_a is initially excited to about 4×10^{-12} m. There is also a small excitation of c_b

(owing to the non-zero overlap of the mechanical resonances); however, this unintentional excitation accounts for less than about 1% of the total energy, and does not qualitatively affect the results presented here.

Comparing $|c_{a,b}(0)|$ with $|c_{a,b}(\tau)|$ in Fig. 3a, b, it is clear that energy is lost from the system during the control loop. This reflects the fact that the damping here is always positive. To distinguish this overall energy loss from effects related to the topological operation, we focus on the relative energy of the two modes before and after the loop.

The data in Fig. 3a were taken for a control loop that did not enclose the EP ($\Delta_{\max} = -1,440$ kHz, $P_{\max} = 750$ μW; for all data, $\Delta_{\min} = -1,890$ kHz, $P_{\min} = 2$ μW). As a result, the nearly adiabatic transit around the control loop results in negligible energy transfer at the end of the control loop. This can be seen qualitatively in Fig. 3a by noting that approximately 99% of the energy is in c_a both immediately before and immediately after the control loop.

By contrast, Fig. 3b shows a measurement in which the control loop does enclose the EP ($\Delta_{\max} = -300$ kHz, $P_{\max} = 750$ μW). The effect on the dynamics is readily visible: before the loop more than 99% of the energy is in c_a , whereas after the loop more than 99% of the (remaining) energy is in c_b .

To quantify the transfer of energy from one mode to another, we define the efficiency $E = |c_b(\tau)|^2 / [|c_a(\tau)|^2 + |c_b(\tau)|^2]$ (this definition makes use of the fact that, before the loop, nearly all the energy is in c_a). The values of $|c_{a,b}(\tau)|$ are determined by fitting decaying exponentials to $|c_{a,b}(t)|$ for $t > \tau + 20$ ms and extrapolating these fits to $t = \tau$.

Figure 3c shows $E(\Delta_{\max})$ for fixed $P_{\max} = 750$ μW; Fig. 3d shows $E(P_{\max})$ for fixed $\Delta_{\max} = -290$ kHz. The limiting behaviour in both cases (that is, for large or small P_{\max} and Δ_{\max}) agrees with the prediction that adiabatic paths enclosing the EP will result in energy transfer, whereas adiabatic paths not enclosing the EP will not. The solid lines in Fig. 3c, d are the results of numerically integrating equations (1) and (2), and are not fits; rather, they use the $P(t)$ and $\Delta(t)$ used in the measurements, and the values of $g_{1,2}$, $\omega_{1,2}$, $\gamma_{1,2}$, κ_{in} and κ determined from the data in Fig. 1. These simulations show good agreement with the measurements irrespective of whether the loop encloses the EP and of whether the loop satisfies adiabaticity.

The measurements shown in Fig. 3 were all made by applying the initial drive to the 'a' mode and then executing a control loop in the counter-clockwise sense. In this case, the adiabatic trajectories enclosing the EP correspond to the less-damped eigenmode (red regions of the surfaces in Fig. 2) for the majority of the loop. By contrast, executing the same loop in the clockwise sense would result in an adiabatic trajectory corresponding primarily to the more-damped eigenmode (blue regions in Fig. 2). As described in refs 12–14, 28, adiabatic behaviour is expected while the system is in the less-damped eigenmode; however, when the system is in the more-damped mode, competition between the non-adiabatic transfer (which is exponentially small in τ) and the effect of differential loss (which is exponentially large in τ) leads to a breakdown of adiabaticity, causing the system to eventually relax to the less-damped mode. This process may also be understood as a consequence of the Stokes phenomenon of asymptotics¹².

This behaviour is demonstrated in Fig. 4, which shows $E(\tau)$ when the EP is encircled in the counter-clockwise or clockwise sense, and with the initial excitation in the 'a' mode (for which E is as defined above) or the 'b' mode (for which E is as defined above, but with the subscripts reversed). The same loop was used in all four cases: $\Delta_{\min} = -1,890$ kHz, $P_{\min} = 2$ μW, $\Delta_{\max} = -290$ kHz and $P_{\max} = 750$ μW. In all four cases, executing the loop very quickly results in negligible energy transfer ($E \rightarrow 0$ as $\tau \rightarrow 0$), consistent with the conventional expectation for a sudden perturbation.

The adiabatic limit ($\tau \gg 1$ ms) is quite different. Efficient energy transfer is achieved ($E \rightarrow 1$) for an initial excitation in the 'a' mode and a counter-clockwise loop (and for an initial excitation in the 'b' mode and a clockwise loop), consistent with the discussion of Fig. 3, and with the fact that these conditions correspond to adiabatic paths almost entirely in the less-damped mode. By contrast, $E \rightarrow 0$ when $\tau \gg 1$ ms

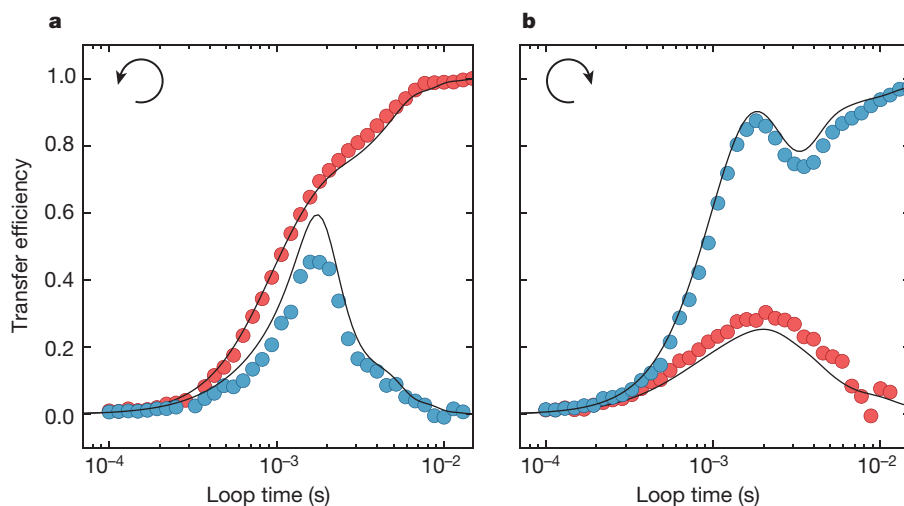


Figure 4 | Non-reciprocal topological dynamics. **a, b,** The transfer efficiency E as a function of the duration of the control loop, τ . The loop shape is identical for all four data series and encloses the EP. The loop is counter-clockwise in **a** and clockwise in **b**, as indicated by the arrows. Red (blue) circles represent data for which the 'a' ('b') mode is initially driven. In all four cases, rapid encircling around the EP ($\tau \rightarrow 0$) results in vanishing energy transfer ($E \rightarrow 0$). For adiabatic encircling, the limiting behaviour of E depends on the sense of the loop and which mode is initially excited. For counter-clockwise (clockwise) loops, the red (blue) data correspond to conventional adiabaticity ($E \rightarrow 1$ as τ increases) and the blue (red) data show the opposite behaviour ($E \rightarrow 0$ as τ increases). As described in the text, this reflects the non-reciprocity of each topological operation (counter-clockwise or clockwise loop). The solid lines are numerical simulations of the dynamics and are completely constrained by the parameters from the fit in Fig. 1.

for an initial excitation in the 'b' mode and a counter-clockwise loop (and for an initial excitation in the 'a' mode and a clockwise loop).

The behaviour described above may be summarized by describing an adiabatic control loop around an EP as a matrix that transforms the initial state $C(0) = [c_1(0), c_2(0)]^T$ to the final state $C(\tau) = [c_1(\tau), c_2(\tau)]^T$ with the form:

$$U_{\odot, \circ}(\tau) = \begin{pmatrix} a_{\odot, \odot}(\tau) & b_{\odot, \odot}(\tau) \\ c_{\odot, \odot}(\tau) & d_{\odot, \odot}(\tau) \end{pmatrix} \quad (4)$$

where \odot and \circ denote a counter-clockwise and clockwise loop, respectively. Because H is a symmetric matrix, it is straightforward to show that if $U_{\odot}(\tau)$ and $U_{\circ}(\tau)$ represent identical but time-reversed control loops, then $U_{\odot} = U_{\circ}^T$. Along with this relationship, the four data-sets in Fig. 4 demonstrate the non-reciprocity of these operations, that is, that $b_{\odot, \odot}(\tau) \neq c_{\odot, \odot}(\tau)$ for $\tau \gg 1$ ms (ref. 29). This inequality is also evident in direct measurements of $|b_{\odot, \odot}(\tau)|$ and $|c_{\odot, \odot}(\tau)|$ (see Methods).

We have demonstrated a new form of adiabatic topological operation that allows for non-reciprocal energy transfer between two eigenmodes of a mechanical system. This transfer exploits the presence of an EP in the spectrum of the two modes. The square membrane used here also offers threefold and fourfold near-degeneracies, opening up the possibility of studying dynamics in the vicinity of higher-order EPs^{15,16}. Furthermore, the cryogenic optomechanical device used here is subject to both thermal and quantum fluctuations³⁰; it is an open question whether non-reciprocal topological effects will allow for new forms of control over these fluctuations.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 January; accepted 11 May 2016.

Published online 25 July 2016.

- Simon, B. Holonomy, the quantum adiabatic theorem, and Berry's phase. *Phys. Rev. Lett.* **51**, 2167–2170 (1983).
- Berry, M. V. Quantal phase factors accompanying adiabatic changes. *Proc. R. Soc. Lond. A* **392**, 45–57 (1984).
- Berry, M. V. Classical adiabatic angles and quantal adiabatic phase. *J. Phys. A* **18**, 15–27 (1985).
- Hannay, J. H. Angle variable holonomy in adiabatic excursion of an integrable Hamiltonian. *J. Phys. A* **18**, 221–230 (1985).
- Arovas, D., Schrieffer, J. R. & Wilczek, F. Fractional statistics and the quantum Hall effect. *Phys. Rev. Lett.* **53**, 722–723 (1984).
- Tomita, A. & Chiao, R. Y. Observation of Berry's topological phase by use of an optical fiber. *Phys. Rev. Lett.* **57**, 937–940 (1986).
- Kitaev, A. Y. Fault-tolerant quantum computation by anyons. *Ann. Phys.* **303**, 2–30 (2003).
- Nayak, C., Simon, S. H., Stern, A., Freedman, M. & Das Sarma, S. Non-Abelian anyons and topological quantum computation. *Rev. Mod. Phys.* **80**, 1083–1159 (2008).

- Heiss, W. D. Phases of wave functions and level repulsion. *Euro. Phys. J. D* **7**, 1–4 (1999).
- Keck, F., Korsch, H. J. & Mossmann, S. Unfolding a diabolic point: a generalized crossing scenario. *J. Phys. A* **36**, 2125–2137 (2003).
- Berry, M. V. Physics of nonhermitian degeneracies. *Czech. J. Phys.* **54**, 1039–1047 (2004).
- Berry, M. V. & Uzdin, R. Slow non-Hermitian cycling: exact solutions and the Stokes phenomenon. *J. Phys. A* **44**, 435303 (2011).
- Uzdin, R., Mailybaev, A. & Moiseyev, N. On the observability and asymmetry of adiabatic state flips generated by exceptional points. *J. Phys. A* **44**, 435302 (2011).
- Milburn, T. J. *et al.* General description of quadiabatic dynamical phenomena near exceptional points. *Phys. Rev. A* **92**, 052124 (2015).
- Cartarius, H., Main, J. & Wunner, G. Exceptional points in the spectra of atoms in external fields. *Phys. Rev. A* **79**, 053408 (2009).
- Demange, G. & Graefe, E.-M. Signatures of three coalescing eigenfunctions. *J. Phys. A* **45**, 025303 (2012).
- Arnold, V. I. *Mathematical Methods of Classical Mechanics* Ch. 10 (Springer, 1989).
- Ando, T., Nakanishi, T. & Saito, R. Berry's phase and absence of back scattering in carbon nanotubes. *J. Phys. Soc. Jpn* **67**, 2857–2862 (1998).
- Lefebvre, R., Atabek, O., Sindelka, M. & Moiseyev, N. Resonance coalescence in molecular photodissociation. *Phys. Rev. Lett.* **103**, 123003 (2009).
- Hamamda, M., Pillet, P., Lignier, H. & Comparat, D. Ro-vibrational cooling of molecules and prospects. *J. Phys. B* **48**, 182001 (2015).
- Kapralová-Žďánská, P. R. & Moiseyev, N. Helium in chirped laser fields as a time-asymmetric atomic switch. *J. Chem. Phys.* **141**, 014307 (2014).
- Kim, S. Braid operation of exceptional points. *Fortschr. Phys.* **61**, 155–161 (2013).
- Philipp, M., von Brentano, P., Pascovici, G. & Richter, A. Frequency and width crossing of two interacting resonances in a microwave cavity. *Phys. Rev. E* **62**, 1922–1926 (2000).
- Dembowski, C. *et al.* Experimental observation of the topological structure of exceptional points. *Phys. Rev. Lett.* **86**, 787–790 (2001).
- Thompson, J. D. *et al.* Strong dispersive coupling of a high-finesse cavity to a micromechanical membrane. *Nature* **452**, 72–75 (2008).
- Aspelmeyer, M., Kippenberg, T. J. & Marquardt, F. Cavity optomechanics. *Rev. Mod. Phys.* **86**, 1391–1452 (2014).
- Jing, H. *et al.* PT -symmetric phonon laser. *Phys. Rev. Lett.* **113**, 053604 (2014).
- Graefe, E.-M., Mailybaev, A. A. & Moiseyev, N. Breakdown of adiabatic transfer of light in waveguides in the presence of absorption. *Phys. Rev. A* **88**, 033842 (2013).
- Jalas, D. *et al.* What is — and what is not — an optical isolator. *Nat. Photon.* **7**, 579–582 (2013).
- Underwood, M. *et al.* Measurement of the motional sidebands of a nanogram-scale oscillator in the quantum regime. *Phys. Rev. A* **92**, 061801 (2015).

Acknowledgements We thank L. Jiang, D. Lee, T. Milburn, P. Rabl, S. Rotter, A. Shkarin and W. Underwood for discussions. This work was supported by AFOSR Grant FA9550-15-1-0270.

Author Contributions H.X., D.M. and L.J. performed the measurements and analysed the data. J.G.E.H. and H.X. wrote the manuscript with input from all the authors. J.G.E.H. directed the research.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.G.E.H. (jack.harris@yale.edu).

METHODS

Measurement set-up. A schematic illustration of the experiment is shown in Extended Data Fig. 1. The optomechanical device and much of the measurement set-up are described in ref. 30. The membrane and optical cavity are mounted in a cryostat that is maintained at $T = 4.2$ K. The motion of the membrane is monitored via a heterodyne measurement using a probe beam and a local oscillator, both produced from a single laser ('ML' in Extended Data Fig. 1a). The probe-beam frequency is shifted by an acousto-optical modulator (AOM1 in Extended Data Fig. 1a) driven at 80 MHz. Pound–Drever–Hall locking is used to keep the probe beam nearly resonant with one mode of the cavity; as a result its detuning $\Delta_p \ll \kappa$, resulting in a negligible contribution to σ . Likewise, the large detuning of the local oscillator ($\Delta_{LO} \approx 80$ MHz $\gg \kappa$) also results in a negligible contribution to σ . Control over the optomechanical system is provided by a separate laser ('CL' in Extended Data Fig. 1a), whose detuning Δ and power P are controlled by an additional acousto-optic modulator (AOM3 in Extended Data Fig. 1a). The frequencies of the various beams are illustrated in Extended Data Fig. 1b. The cavity is approximately single-sided and all measurements are performed in reflection. The reflected beams are incident on a single photodiode, and demodulation circuits are used to monitor multiple Fourier components of the heterodyne signal, each with a bandwidth equal to 50 Hz.

Optically mediated mechanical coupling. We consider a system consisting of two mechanical modes, each coupled linearly to a common optical mode. We show that the optical field generates a tunable effective coupling between the mechanical modes, which can be exploited to produce an EP, as described in the main text. The model closely follows the one presented in ref. 31.

In a standard optomechanical system, one considers an optical cavity mode with a frequency that is linearly coupled to the position of a mechanical oscillator. An input–output approach to this system yields a pair of coupled differential equations for the two modes, which can be easily treated in the Fourier domain to understand the optical modification of the mechanical susceptibility. Here we consider a simple extension of this model in which there are two mechanical modes, each coupled to the same optical mode. This yields the following equations of motion for the mechanical/optical modes:

$$\begin{aligned}\dot{a} &= -\left(\frac{\kappa}{2} + i\omega_c\right)a - ig_1az_1 - ig_2az_2 + \sqrt{\kappa_{\text{in}}}a_{\text{in}} \\ \dot{c}_1 &= -\left(\frac{\gamma_1}{2} + i\omega_1\right)c_1 - ig_1a^*a + \sqrt{\gamma_1}\eta_1 \\ \dot{c}_2 &= -\left(\frac{\gamma_2}{2} + i\omega_2\right)c_2 - ig_2a^*a + \sqrt{\gamma_2}\eta_2\end{aligned}$$

where a is the optical mode amplitude with resonant frequency ω_c , total dissipation rate κ and input coupling rate κ_{in} . The i th mechanical mode is described by position $z_i = c_i + c_i^*$, where c_i is the complex mode amplitude and the asterisks indicate complex conjugation. Each mechanical mode has resonant frequency ω_i , dissipation rate γ_i , and is coupled to the optical mode with a single-photon coupling rate g_i . The optical and mechanical modes are driven by input fields a_{in} and η_i , respectively.

We now suppose that the cavity is driven by a beam with power P and frequency Ω_L , detuned from the cavity resonance by $\Delta = \Omega_L - \omega_c$. By doing so, we can express the optical field as fluctuations $d(t)$ around a mean intracavity field given by

$$\bar{a} = \frac{\sqrt{\kappa_{\text{in}}}}{\frac{\kappa}{2} - i\Delta} a_{\text{in}}; \quad a_{\text{in}} = \sqrt{\frac{P}{\hbar\Omega_L}}$$

Making these substitutions in the original system of equations yields the linearized equations of motion:

$$\begin{aligned}\dot{d} &= -\left(\frac{\kappa}{2} - i\Delta\right)d - i\alpha_1z_1 - i\alpha_2z_2 \\ \dot{c}_1 &= -\left(\frac{\gamma_1}{2} + i\omega_1\right)c_1 - i(\alpha_1^*d + \alpha_1d^*) + \sqrt{\gamma_1}\eta_1 \\ \dot{c}_2 &= -\left(\frac{\gamma_2}{2} + i\omega_2\right)c_2 - i(\alpha_2^*d + \alpha_2d^*) + \sqrt{\gamma_2}\eta_2\end{aligned}$$

where we have defined $\alpha_i = \bar{a}g_i$. Moving to the Fourier domain, and defining the cavity susceptibility $\chi_c(\omega) = [\kappa/2 - i(\omega + \Delta)]^{-1}$, we solve for $d(\omega)$ and $d^*(\omega)$ and substitute these into the equations for $c_{1,2}(\omega)$ to find a reduced system of two equations describing the mechanical modes:

$$\begin{aligned}\left[\frac{\gamma_1}{2} - i(\omega - \omega_1)\right]c_1(\omega) &= |\alpha_1|^2 [\chi_c^*(-\omega) - \chi_c(\omega)]c_1(\omega) \\ &\quad + \alpha_1^*\alpha_2 [\chi_c^*(-\omega) - \chi_c(\omega)]c_2(\omega) \\ \left[\frac{\gamma_2}{2} - i(\omega - \omega_2)\right]c_2(\omega) &= |\alpha_2|^2 [\chi_c^*(-\omega) - \chi_c(\omega)]c_2(\omega) \\ &\quad + \alpha_1^*\alpha_2 [\chi_c^*(-\omega) - \chi_c(\omega)]c_1(\omega)\end{aligned}$$

Note that we have dropped counter-rotating c_1^* and c_2^* terms. We have also dropped the mechanical drive terms $\eta_{1,2}$. These are not necessary for our model, because we drive the system to a particular initial state, turn off the drive and then focus on the evolution of the system without any mechanical drive applied.

In the traditional optomechanical system, one defines the (single-mode) optomechanical self-energy as $\Sigma_{\text{SM}}(\omega) = i|\alpha|^2 [\chi_c^*(-\omega) - \chi_c(\omega)]$. In this two-mode system, we can extend this concept to a self-energy matrix:

$$\Sigma = \begin{pmatrix} i|\alpha_1\alpha_1| & i|\alpha_1\alpha_2| \\ i|\alpha_2\alpha_1| & i|\alpha_2\alpha_2| \end{pmatrix} [\chi_c^*(-\omega) - \chi_c(\omega)] = \begin{pmatrix} -ig_1^2\sigma & -ig_1g_2\sigma \\ -ig_1g_2\sigma & -ig_2^2\sigma \end{pmatrix} \quad (5)$$

where σ is defined in equation (3).

Writing our mechanical modes as a vector $C(t) = [c_1(t), c_2(t)]^T$, we can write the following matrix equation:

$$-i\omega C(\omega) = -\begin{pmatrix} \frac{\gamma_1}{2} + i\omega_1 & 0 \\ 0 & \frac{\gamma_2}{2} + i\omega_2 \end{pmatrix} C(\omega) - i\Sigma(\omega)C(\omega)$$

Before we move back to the time domain, we note that $\Sigma(\omega)$ varies on the scale of κ , whereas the mechanical modes are susceptible to drives only within their linewidth, which is substantially smaller than κ , by assumption. Therefore, it is sufficient to consider $\Sigma(\omega) \approx \Sigma(\omega_1) \approx \Sigma(\omega_2) \equiv \Sigma$. (The mechanical modes are also assumed to be nearly degenerate.) Now that Σ is not a function of ω , we can easily move back to the time domain to obtain equation (1) (reprinted here for convenience):

$$i\dot{C}(t) = HC(t)$$

where we define

$$H = \begin{pmatrix} \omega_1 - i\frac{\gamma_1}{2} & 0 \\ 0 & \omega_2 - i\frac{\gamma_2}{2} \end{pmatrix} + \Sigma \quad (6)$$

Here Σ is a complex quantity, which depends (via α_1 and α_2) on P and Δ . This is the tunability that allows us to access an EP in the spectrum of the two mechanical modes.

We note that equation (6) is identical to equation (2); the apparent difference is due to the fact that in equation (2) the matrix Σ is expressed using the right-most form in equation (5).

Measuring the mechanical eigenvalue spectrum. In Figs 1 and 2, we show the presence of an EP in the complex eigenvalue spectrum (frequencies and decay rates) of the mechanical modes. At each point (P, Δ) , the eigenvalues were measured by optically driving the mechanical modes and measuring their driven response. We measure the mechanical sidebands using the heterodyne measurement laser, locked to the cavity resonance. We set a certain P and Δ for the control laser, then apply amplitude modulation at a frequency near ω_1 and ω_2 , thus creating an optical beat note that drives the mechanical modes. This modulation frequency is swept over ω_1 and ω_2 , and we use a lock-in amplifier to measure the complex response of the heterodyne signal to this drive.

Two examples of these measurements are shown here. Extended Data Fig. 2 shows a sweep over the two modes when the control-beam power is low and there is minimal hybridization of the two modes. In Extended Data Fig. 3, the control-beam power is large and detuned near $-\omega_{1,2}$ such that the modes hybridize substantially, resulting in modes with nearly degenerate frequencies, but different linewidths. The relative phase of the driven response of the two modes is such that we see destructive interference in Extended Data Fig. 3. By fitting the complex response to a sum of complex Lorentzians with an arbitrary phase offset, we extract $\omega_1, \omega_2, \gamma_1$ and γ_2 . The solid lines in Extended Data Figs 2 and 3 are these fits, from which we extract the eigenvalues plotted in Figs 1 and 2.

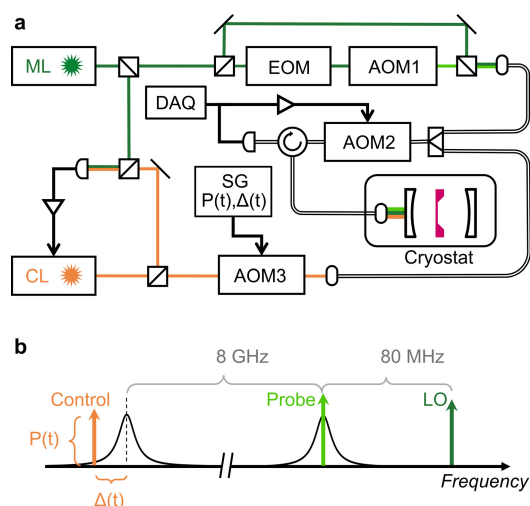
EPs have also been observed in many other systems, including atom–cavity composites³², microwave cavities^{23,24}, optical systems^{33–36}, electronic circuits³⁷ and an exciton–polariton system³⁸, and are predicted to exist in Bose–Einstein condensates^{39,40}, quantum dots⁴¹, acoustic systems⁴², magnetohydrodynamic dynamos⁴³ and nuclei⁴⁴.

Measurement of propagator matrix elements. Figure 4 shows the non-reciprocity of the topological operations as parameterized by their energy transfer efficiency E . The non-reciprocity of these operations can also be seen from direct measurements of the magnitudes of the matrix elements defined in equation (4). These measurements are carried out by, for example, initially driving the ‘a’ mode and then performing a clockwise loop about the EP; in this case $|a_{\odot}(\tau)| = |c_a(\tau)/c_a(0)|$ and $|c_{\odot}(\tau)| = |c_b(\tau)/c_a(0)|$. Similarly, repeating this process, but with the ‘b’ mode initially driven, gives $|b_{\odot}(\tau)|$ and $|d_{\odot}(\tau)|$. In Extended Data Fig. 4, we plot the magnitudes of these propagator matrix elements as a function of the loop duration τ . The points in Extended Data Fig. 4 are extracted from the same data as shown in Fig. 4. For sufficiently large τ , we see that $|b_{\odot, \odot}(\tau)| \neq |c_{\odot, \odot}(\tau)|$, as stated in the main text, which implies $U_{\odot, \odot}(\tau) \neq U_{\odot, \odot}^T(\tau)$.

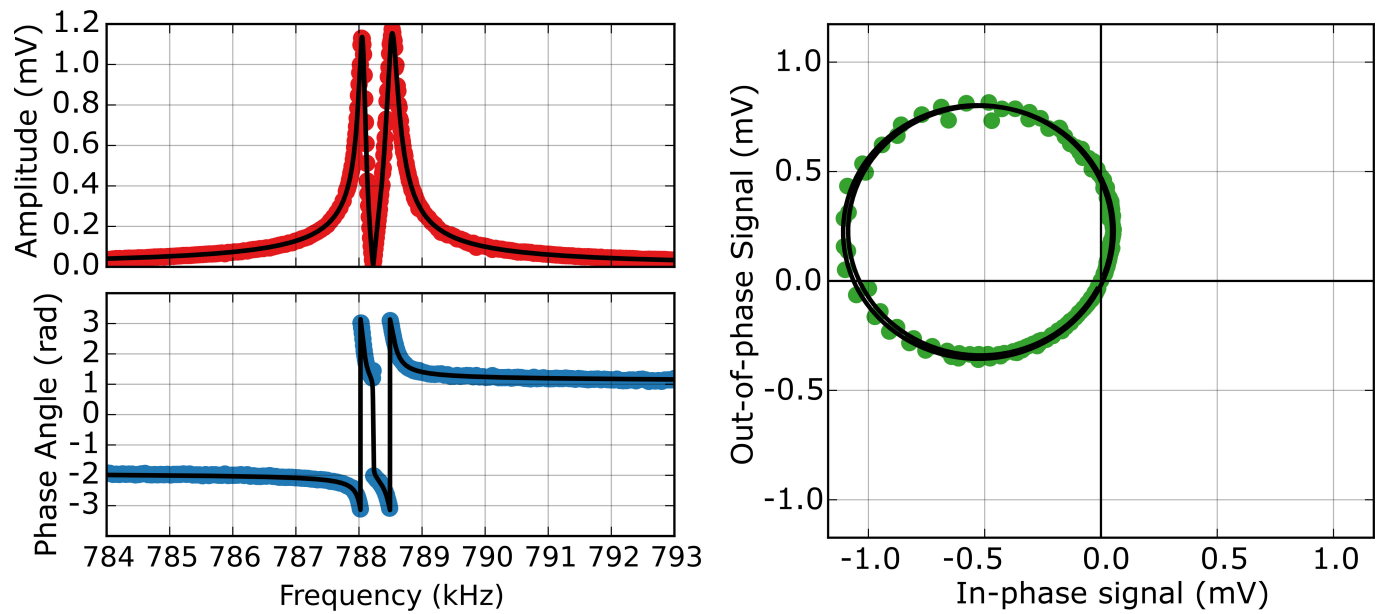
The real-time dynamics studied here can be connected to the propagation of light through an optical crystal with properties that vary along the beam path^{12,45}. An encircling around an EP is also mapped onto the propagation through a two-mode waveguide in a concurrent experiment⁴⁶.

31. Shkarin, A. B. *et al.* Optically mediated hybridization between two mechanical modes. *Phys. Rev. Lett.* **112**, 013602 (2014).
32. Choi, Y. *et al.* Quasieigenstate coalescence in an atom-cavity quantum composite. *Phys. Rev. Lett.* **104**, 153601 (2010).

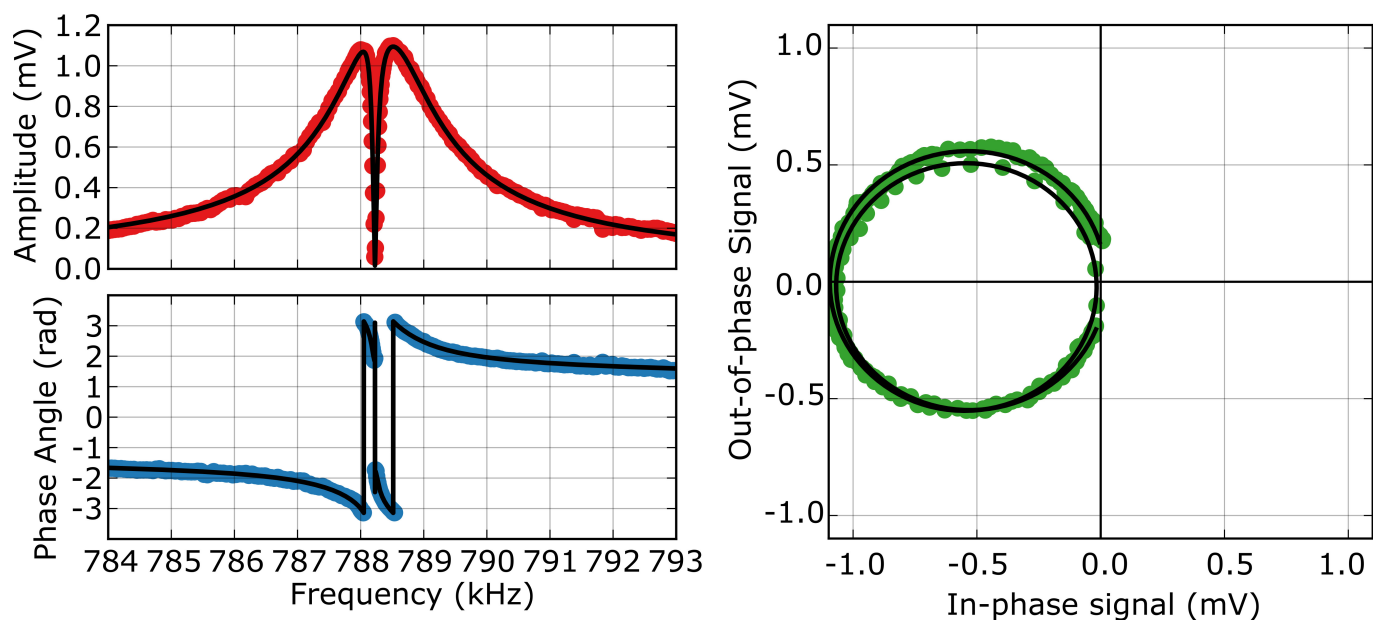
33. Lee, S.-B. *et al.* Observation of an exceptional point in a chaotic optical microcavity. *Phys. Rev. Lett.* **103**, 134101 (2009).
34. Brandstetter, M. *et al.* Reversing the pump dependence of a laser at an exceptional point. *Nat. Commun.* **5**, 4034 (2014).
35. Peng, B. *et al.* Loss-induced suppression and revival of lasing. *Science* **346**, 328–332 (2014).
36. Zhen, B. *et al.* Spawning rings of exceptional points out of Dirac cones. *Nature* **525**, 354–358 (2015).
37. Stehmann, T., Heiss, W. D. & Scholtz, F. G. Observation of exceptional points in electronic circuits. *J. Phys. A* **37**, 7813–7819 (2004).
38. Gao, T. *et al.* Observation of non-Hermitian degeneracies in a chaotic exciton-polariton billiard. *Nature* **526**, 554–558 (2015).
39. Heiss, W. D. & Nazmitdinov, R. G. Instabilities, nonhermiticity and exceptional points in the cranking model. *J. Phys. A* **40**, 9475–9481 (2007).
40. Cartarius, H., Main, J. & Wunner, G. Discovery of exceptional points in the Bose-Einstein condensation of gases with attractive $1/r$ interaction. *Phys. Rev. A* **77**, 013618 (2008).
41. Weidenmüller, H. A. Crossing of two Coulomb blockade resonances. *Phys. Rev. B* **68**, 125326 (2003).
42. Wu, T.-T. & Huang, Z.-G. Level repulsions of bulk acoustic waves in composite materials. *Phys. Rev. B* **70**, 214304 (2004).
43. Günther, U., Stefani, F. & Gerbeth, G. The MHD α^2 -dynamo, Z_2 -graded pseudo-Hermiticity, level crossings and exceptional points of branching type. *Czech. J. Phys.* **54**, 1075–1089 (2004).
44. Michel, N., Nazarewicz, W., Okołowicz, J. & Płoszajczak, M. Open problems in the theory of nuclear open quantum systems. *J. Phys. G* **37**, 064042 (2010).
45. Berry, M. V. Optical polarization evolution near a non-Hermitian degeneracy. *J. Opt.* **13**, 115701 (2011).
46. Doppler, J. *et al.* Dynamically encircling an exceptional point for asymmetric mode switching. *Nature* <http://www.doi.org/10.1038/nature18605> (2016).



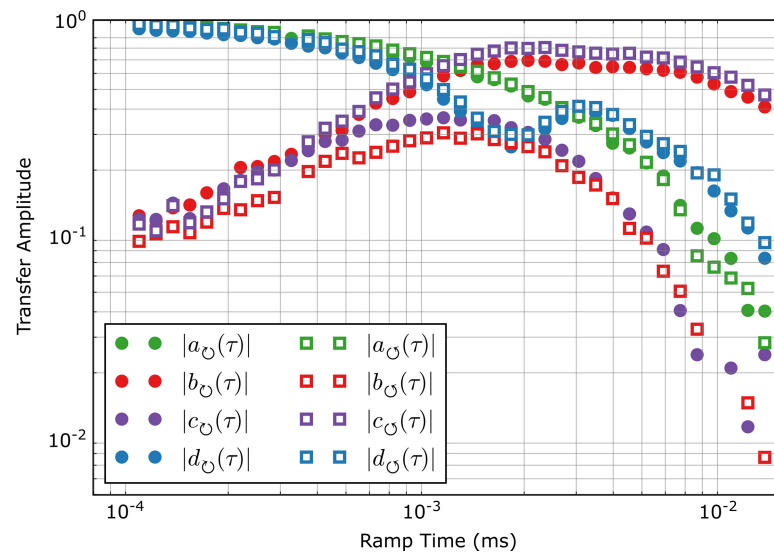
Extended Data Figure 1 | Experimental schematics. **a**, Illustration of the optical and electronic components. The measurement laser ('ML') is split into a local oscillator ('LO' in **b**) and a probe beam ('Probe' in **b**). The probe-beam frequency is shifted by an acousto-optic modulator ('AOM1'), and is locked to the cavity using a Pound-Drever-Hall (PDH) scheme and modulation produced by an electro-optic modulator ('EOM'). The control laser ('CL'; 'Control' in **b**) is locked to the measurement laser with a frequency offset that is approximately double the free spectral range of the cavity. The control parameters used to access the EP are the power P and detuning Δ of the control laser. P and Δ are set by the amplitude and frequency of a signal generator ('SG'), which drives another acousto-optic modulator ('AOM3'). The PDH error signal is used to control the frequency of yet another acousto-optic modulator ('AOM2'), ensuring that all beams track fluctuations of the cavity. Light is delivered to (and collected from) the cryostat via an optical circulator. Coloured lines, hollow lines and thick black lines show free-space laser beams, optical fibres and electrical circuits, respectively. Triangles, ovals and semicircles show electronics, fibre couplers and photodiodes, respectively. 'DAQ' indicates the data acquisition system. The silicon nitride membrane is shown in purple. **b**, Illustration of the optical frequency domain. Lasers are indicated by coloured arrows and cavity modes by black curves.



Extended Data Figure 2 | Lock-in signal at low laser power ($\Delta = -780$ kHz, $P = 73$ μ W). Left, amplitude (top, red) and phase angle (bottom, blue) of the lock-in signal as a function of drive frequency. Right, the same data shown as a parametric plot of the in-phase and out-of-phase components of the lock-in signal as a function of drive frequency.



Extended Data Figure 3 | Lock-in signal at high laser power ($\Delta = -780$ kHz, $P = 380 \mu\text{W}$). Left, amplitude (top, red) and phase angle (bottom, blue) of the lock-in signal as a function of drive frequency. Right, the same data shown as a parametric plot of the in-phase and out-of-phase components of the lock-in signal as a function of drive frequency.



Extended Data Figure 4 | Magnitudes of propagator matrix elements.

Ablation-cooled material removal with ultrafast bursts of pulses

Can Kerse¹, Hamit Kalaycıoğlu², Parviz Elahi², Barbaros Çetin³, Denizhan K. Kesim¹, Önder Akçaalan¹, Seydi Yavaş⁴, Mehmet D. Aşik⁵, Bülent Öktem⁶, Heinar Hoogland^{7,8}, Ronald Holzwarth⁷ & Fatih Ömer Ilday^{1,2}

The use of femtosecond laser pulses allows precise and thermal-damage-free removal of material (ablation) with wide-ranging scientific^{1–5}, medical^{6–11} and industrial applications¹². However, its potential is limited by the low speeds at which material can be removed^{1,9–11,13} and the complexity of the associated laser technology. The complexity of the laser design arises from the need to overcome the high pulse energy threshold for efficient ablation. However, the use of more powerful lasers to increase the ablation rate results in unwanted effects such as shielding, saturation and collateral damage from heat accumulation at higher laser powers^{6,13,14}. Here we circumvent this limitation by exploiting ablation cooling, in analogy to a technique routinely used in aerospace engineering^{15,16}. We apply ultrafast successions (bursts) of laser pulses to ablate the target material before the residual heat deposited by previous pulses diffuses away from the processing region. Proof-of-principle experiments on various substrates demonstrate that extremely high repetition rates, which make ablation cooling possible, reduce the laser pulse energies needed for ablation and increase the efficiency of the removal process by an order of magnitude over previously used laser parameters^{17,18}. We also demonstrate the removal of brain tissue at two cubic millimetres per minute and dentine at three cubic millimetres per minute without any thermal damage to the bulk^{9,11}.

Ablation is the evaporative removal of a material when its temperature exceeds a critical value. Because the ablated material is physically carried away, the thermal energy contained in the ablated mass is also removed, thus reducing the average temperature of the remaining material. This effect forms the basis of ablation cooling, which has been routinely used as an approach to thermal protection during the atmospheric re-entry of rockets since the 1950s, owing to the minimal mass requirements¹⁵. Unlike ablation cooling for rockets, laser ablation is not continuous, but takes place only during and shortly after an incident laser pulse. For the laser parameters used in previous experiments ablation cooling has been negligible as a cooling mechanism in comparison with heat conduction (diffusion) from the processing region into the bulk of the target, which is continuously occurring. For ablation cooling to become a major contributor, the time delay between the laser pulses (the inverse of the repetition rate) must be reduced until the part of the material that is to be ablated does not cool substantially between successive pulses. Only then would heat extraction due to ablation become comparable to that due to diffusion (Fig. 1a).

The physics of the ablation-cooled regime can be explained through a toy model (see Supplementary Information section 1 for full details). We assume that each pulse gives rise to an instantaneous temperature rise of ΔT , which is roughly proportional to the pulse energy, E_p , and that the material cools with a $1/\sqrt{1+t/\tau_0}$ dependence on the time delay, t , after the arrival of a pulse. The thermal relaxation

time, τ_0 , is proportional to δ^2/α , where δ is the depth or the lateral radius (whichever dimension is smaller) of the section of the material to be ablated and α is its thermal diffusivity. For a train of N pulses, the temperature of the target surface that is encountered by the $(n+1)$ th pulse is given by $T_{n+1} = T_n + \delta T$, where $\delta T = \Delta T / \sqrt{1 + \tau_R/\tau_0}$ is the small net increase in target temperature by a single pulse and τ_R is inverse of the repetition rate. Ablation occurs when the temperature exceeds a critical value T_c . For the traditional regime of ultrafast ablation, the repetition rate is low ($\tau_R \gg \tau_0$) and each pulse must be energetic enough to cause ablation ($\Delta T > T_c - T_0$, where T_0 is the initial surface temperature). The ablation-cooled regime corresponds to $\tau_R \lesssim \tau_0$. In this regime, the energy of the individual pulses can be lower than the ablation threshold because temperature builds up from pulse to pulse and ablation starts after the m th pulse in the train, where $m = (T_c - T_0 - \Delta T + \delta T)/\delta T$. The volume of the ablated material is given by $V_{\text{ablated}} = \beta[N - u(T_c - T_0 - \Delta T)m]E_p u(N - m)$, where β is a proportionality factor and u is the Heaviside (unit step) function. The thermal energy that diffuses into the bulk of the target owing to cooling between the pulses is

$$E_{\text{heat}} = \alpha(T_c - T_0) \left(1 - \frac{1}{\sqrt{1 + \tau_R/\tau_0}} \right) (N - m)E_p + \alpha(\Delta T - \delta T)mE_p.$$

For the traditional regime, this result reduces to $\lim_{\tau_R \rightarrow \infty} E_{\text{heat}} = \alpha(T_c - T_0)NE_p$.

The toy model makes two main predictions for the ablation-cooled regime—both are confirmed by numerical solutions of the heat diffusion equation (see Supplementary Information section 2 for details) as well as the experiments described below. The first is that increasing the repetition rate reduces the heating of surrounding regions (Fig. 1b, c and Supplementary Fig. 1). Because less of the deposited laser energy is lost to heat diffusion ($\lim_{\tau_R \rightarrow 0} E_{\text{heat}} = 0$), the ablation effi-

ciency is higher than for the traditional regime (Supplementary Fig. 3). The second states that the pulse energy can be decreased if the number of pulses is simultaneously increased in proportion, without a subsequent reduction in the ablation efficiency (Fig. 1d). This is necessary to fully benefit from the ablation-cooled regime, because shielding effects (that is, ablation-induced plasma and ejected particulates reflecting and scattering incoming light) will prevent efficient ablation if the repetition rate is increased at a constant energy¹⁸.

To demonstrate ablation cooling, a customized femtosecond fibre laser^{19–21} was used (see Supplementary Information section 3 for details). We implemented burst-mode operation²², because continuous trains of energetic pulses at the high repetition rates required to access the ablation-cooled regime correspond to a prohibitively high average power and laser repositioning in continuous mode is limited. In burst mode, the laser produces groups of high-repetition-rate pulses, which are, in turn, repeated with a lower frequency. The duty cycle of the

¹Department of Electrical and Electronics Engineering, Bilkent University, Ankara 06800, Turkey. ²Department of Physics, Bilkent University, Ankara 06800, Turkey. ³Department of Mechanical Engineering, Bilkent University, Ankara 06800, Turkey. ⁴FiberLAST, Inc., Ankara 06531, Turkey. ⁵Nanotechnology and Nanomedicine Department, Hacettepe University, Ankara 06800, Turkey. ⁶ASELSAN, Ankara 06150, Turkey. ⁷Menlo Systems GmbH, Am Kloperspitz 19a, Martinsried 82152, Germany. ⁸Lehrstuhl für Laserphysik, Department Physik, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen 91058, Germany.

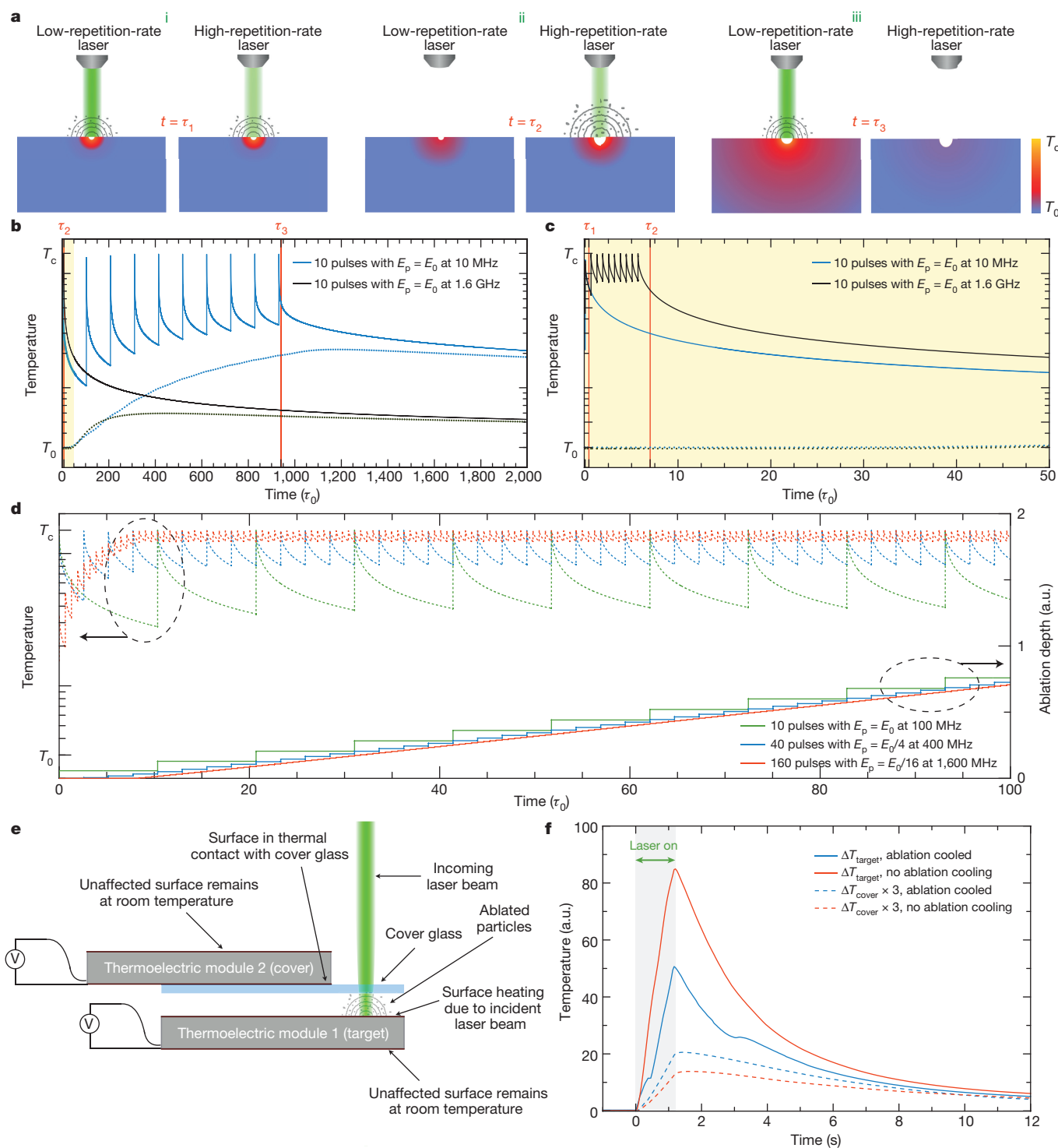


Figure 1 | Principles of ablation-cooled removal of a material by laser.

a, Schematic representation of the ablation process for low (traditional regime, left diagrams) and high (ablation-cooled regime, right diagrams) repetition rates. Temperature profiles are illustrated for $t = \tau_1$ (i), which is shortly after the arrival of the first pulse for both cases; for $t = \tau_2$ (ii), which is before (shortly after) the arrival of the second (last) pulse for the low-repetition-rate (high-repetition-rate) laser; and for $t = \tau_3$ (iii), which is shortly after the arrival of the last pulse for the low-repetition-rate laser. The colouration of the target material is based on simulation results shown in **b** at the indicated time intervals of τ_1 , τ_2 and τ_3 . **b**, Calculated evolution of the temperatures at the surface (solid lines) and below (at a depth of 30 times the optical penetration depth) the surface (dotted lines) for repetition rates of 10 MHz (blue lines) and 1,600 MHz (black lines). The pulse energies and number of pulses are the same for both cases. The higher repetition rate results in substantially lower temperatures below

the surface due to ablation cooling. **c**, Expanded view of the shaded section of the plot in **b**. **d**, Calculated evolution of the surface temperature (dashed lines) and amount of ablated material (solid lines) for repetition rates of 100 MHz (green lines), 400 MHz (blue lines) and 1,600 MHz (red lines). The ablation rate remains approximately the same when the product of the pulse energy and repetition rate is maintained. The spikes in the surface temperatures precisely indicate the arrival of pulses, which are not shown explicitly for clarity. **e**, Experimental set-up for direct confirmation of the ablation-cooling effect. **f**, The measured temperature increase that is induced on thermoelectric module 1 (the target material; solid lines) and thermoelectric module 2 (attached to the coverslip that collects a portion of the ablated particles; dashed lines, values have been multiplied by three to aid comparison with ΔT_{target}) with the laser operating in the ablation-cooled regime (blue lines) and in the traditional regime (red lines).

pulsation can be adjusted to set the average power. Burst-mode material processing has substantial benefits^{22–24}, but the possibility of ablation cooling has not yet been recognized.

First we present experimental evidence of the ablation-cooling effect by simultaneously measuring the temperature of a target material directly and the heat carried by the ablated particles (indirectly) (Fig. 1e). The laser beam is focused onto and ablates the surface of a thermoelectric module. This causes a temperature difference between the laser-targeted top surface and the bottom surface, which generates a voltage difference by the Seebeck effect. A portion of the particles ejected from the surface during ablation stick to a glass coverslip, which is held approximately 1 mm above the target. A second thermoelectric module is used to monitor the temperature of the coverslip, which rises in proportion to the thermal energy delivered by the ablated particles. The measured temperatures of the target and the coverslip (Fig. 1f) confirm that the target heats less, and the coverslip more, in the ablation-cooled regime. The laser parameters were 50 pulses of 3 μJ each, with an 800 fs duration for a 0.2 MHz burst and a 1.7 GHz intraburst repetition rate. This is within the ablation-cooled regime assuming a typical thermal diffusivity of about 150 mm^2s^{-1} for the ceramic surface of the thermoelectric module and 3 μJ , 800 fs pulses at a 10 MHz uniform repetition rate to illustrate the traditional regime (10 MHz was chosen to be safely outside the ablation-cooled regime, although the thermal diffusivity of the ceramic surface is not precisely known).

We demonstrate validity of the predictions of the toy model for ablation cooling across a range of materials (see Supplementary Information for a discussion of other materials). Copper and silicon were chosen as

examples of metal and semiconductor targets, respectively, because their ablation rates with ultrafast pulses are well documented. The volume of material ablated as a function of the incident energy is shown in Fig. 2a for Cu and Fig. 2b for Si for various repetition rates. Figure 2c, d shows the number of atoms ablated per incident photon as a function of the pulse energy. We observe a substantial increase in ablation when the repetition rate is about 100 MHz or higher. Although it is not possible to predict the precise frequency required for each material (the toy model is too simple for us to expect quantitatively accurate predictions), $\tau_0 \approx 1$ ns for Cu for a processing region depth of a few hundred nanometres. Given that increases in efficiency are predicted to begin at a tenth of the corresponding repetition rate, this prediction agrees with the experimental observations. The lower thermal diffusivity of Si compared with Cu is consistent with the increase in its ablation efficiency at 27 MHz, whereas the results at 1 MHz and 27 MHz are similar for Cu, implying that the onset of the ablation-cooled regime for Cu begins between 27 MHz and 108 MHz. If the repetition rate is further increased, efficiency saturates at high pulse energies—a consequence of the expected shielding effect. The solution is to decrease the pulse energy, and increase the number of pulses and the repetition rate (for example, from 25 pulses at 108 MHz to 800 pulses (with 32 times lower energy) at 3,464 MHz). The amount of ablation remains similar (black and pink data in Fig. 2a, b), which means that the shielding effects have been overcome.

To place the ablation results into context, they should be compared with common literature values (see Supplementary Information section 5 for an extensive discussion). Comparison with experiments on Cu using 70 fs pulses with a pulse energy of up to 0.4 mJ at 800 nm (ref. 17)

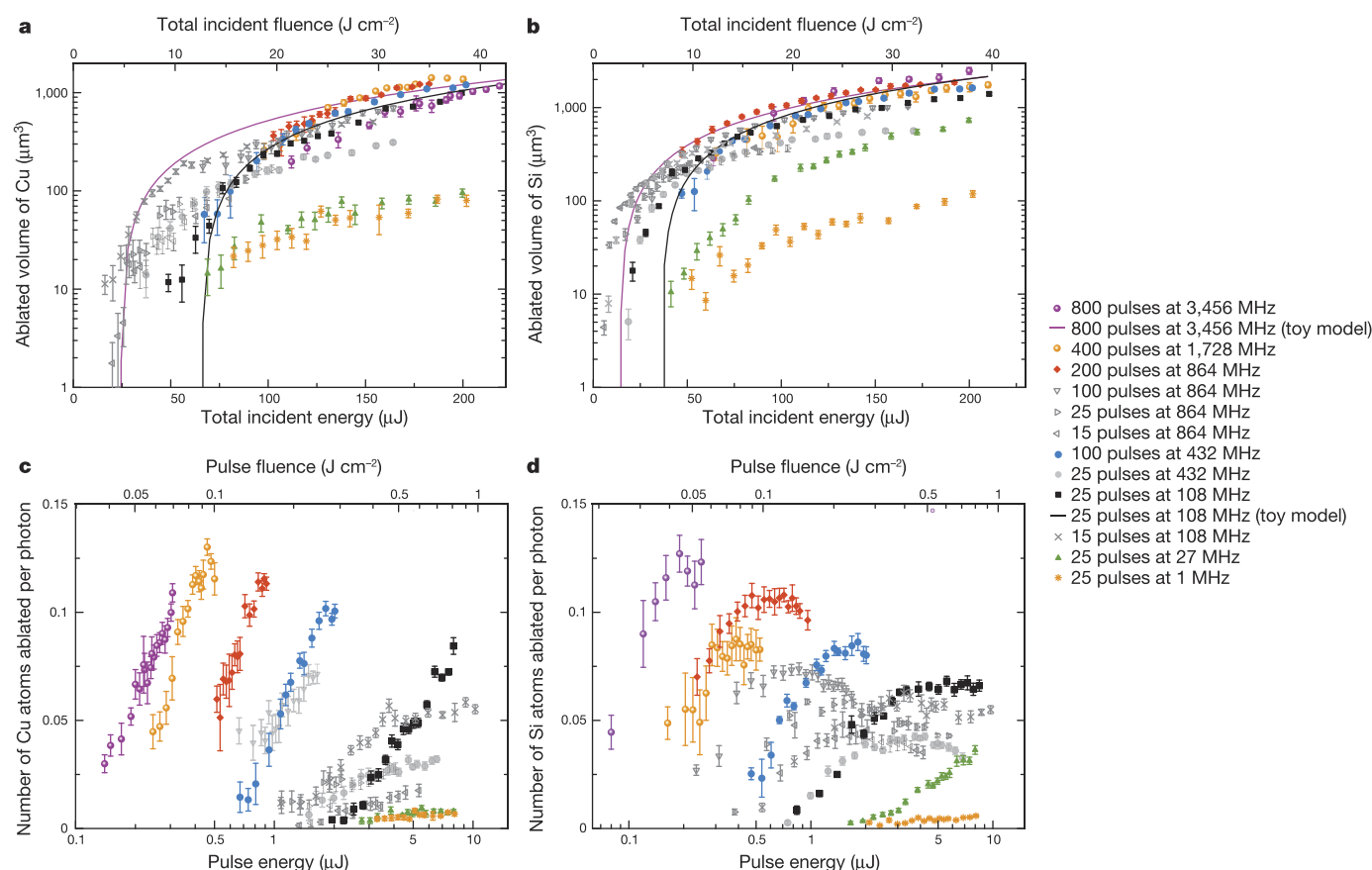


Figure 2 | Scaling down of the pulse energy with increasing repetition rate. **a, b**, Volumes (symbols) of Cu (**a**) and Si (**b**) ablated by a single burst of pulses as a function of total incident energy and fluence for different intraburst repetition rates. The predictions of the toy model for the lowest and highest repetition rates in the ablation-cooled regime are also shown (solid lines). **c, d**, Ablation efficiency in terms of number of atoms of Cu (**c**) and Si (**d**) ablated per incident photon as a function of pulse energy and

pulse fluence for different repetition rates. The legend applies to all panels. The lower and upper limits to the data correspond to the ablation threshold and available laser energy, respectively. In all panels the sample size for each data point is 20, where the centre values represent the mean and the error bars represent the standard deviation. Coloured symbols highlight the onset of the ablation-cooled regime and (beyond 108 MHz) the inverse scaling of the pulse energy with repetition rate in the ablation-cooled regime.

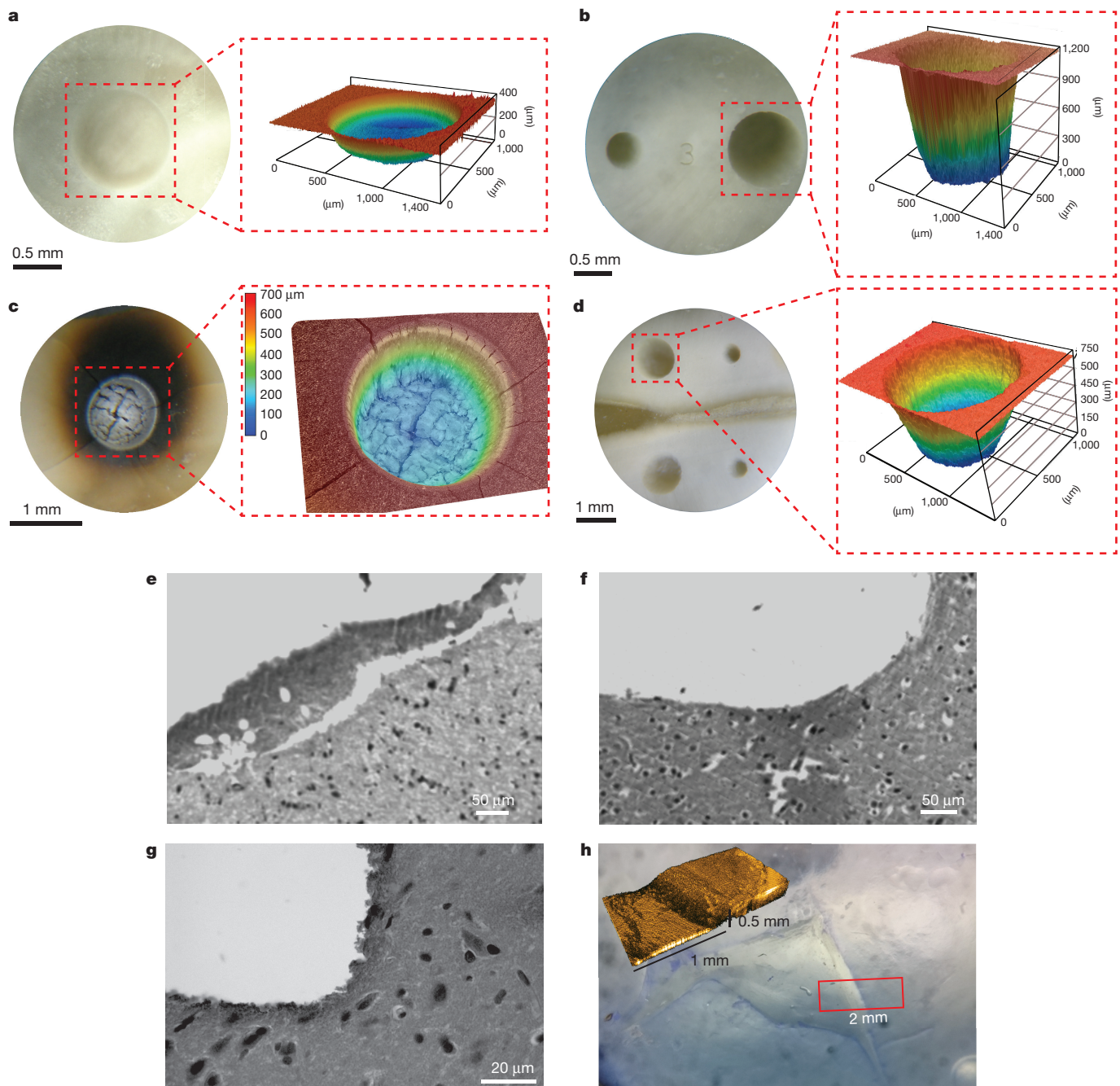


Figure 3 | Ablation of hard and soft tissues. **a, b**, Laser removal of a section of human dentine obtained in the traditional regime (**a**, 1 kHz uniform repetition rate) and in the ablation-cooled regime (**b**, 1.7 GHz intraburst repetition rate). Although both ablation cooling and traditional ultrafast processing avoid thermal damage at sufficiently low average powers, the ablation-cooled regime achieves approximately six times more ablation despite using pulse energies that are about 12 times lower. **c, d**, When the (uniform or intraburst, respectively) repetition rate, average power and scanning speed are simultaneously increased by a factor of 25, the traditional regime of ultrafast processing results in thermal damage (**c**; Supplementary Video 4), whereas the ablation-cooled regime completely avoids thermal effects and achieves an ablation speed of $3 \text{ mm}^3 \text{ min}^{-1}$, despite using a pulse

reveals that we obtain around 2,000 times more ablation at the same fluence of approximately 0.04 J cm^{-2} , which is our maximum fluence for the intraburst repetition rate of 3,456 MHz. Even if we imagine the entire burst of 800 pulses to act like a single pulse and compare the results with those in ref. 17 for an equal total fluence (20 J cm^{-2} , the highest value for which direct comparison is possible), we obtain about 12 times more ablation, although our pulse fluence (energy) is

energy that is 25 times lower (**d**; Supplementary Video 5). The insets in **a–d** show laser scanning microscope characterizations of the ablated holes. **e, f**, Histological images corresponding to about 1 mm^3 sections, which were removed from a rat brain with the laser operating at an average power of 600 mW in the traditional regime (**e**), showing presence of thermal damage, and in the ablation-cooled regime (**f**), showing no major thermal damage. **g**, Ablation-cooled laser removal of brain tissue at an average power of 2.7 W, achieving an ablation speed of $2 \text{ mm}^3 \text{ min}^{-1}$ and showing no major thermal damage. **h**, Bright-field optical image of a bovine cornea from which a flap was removed following ablation-cooled laser processing of a section 0.4 mm below the surface. Inset, optical coherence tomography image of the section indicated by the rectangle.

smaller by a factor of 800 (2,400). Comparison with another reference¹⁸ indicates that the efficiency of ablation in our experiments is 100 times higher despite using a pulse energy that is 260 times lower, when matching the fluence of the entire burst to that of the single-pulse fluence. We achieve a level of ablation that is five times higher than results obtained with a burst-mode laser²³ that does not exploit ablation cooling, despite using a pulse fluence that is 165 times smaller for the

same burst fluence of 20 J cm^{-2} . These results conclusively demonstrate that the exploitation of ablation cooling increases the ablation efficiency by an order of magnitude while allowing the required pulse energy to be reduced by three orders of magnitude.

We now focus on the reduction of undesired thermal effects in the ablation-cooled regime. We have performed systematic comparisons using high and low repetition rates of the same laser with identical focusing and scanning systems. Tissue removal may well be regarded as the ultimate test of the suppression of thermal effects because an increase in temperature of only a few degrees can lead to degradation. Hard-tissue experiments were conducted on human dentine to contrast the ablation-cooled regime with the traditional regime. At low average powers, the traditional regime (using $100 \mu\text{J}$ pulses at 1 kHz , Fig. 3a) and the ablation-cooled regime (25 pulses of $4 \mu\text{J}$ energy at a 1.7 GHz intraburst repetition rate and 1 kHz burst repetition rate, Fig. 3b) both provide results with negligible thermal damage (although the latter achieves an ablation rate four times higher). When increasing the processing speed by a factor of 25 with a corresponding increase in power, the traditional regime causes excessive carbonization (Fig. 3c), whereas the ablation-cooled regime does not, while achieving an ablation rate of $3 \text{ mm}^3 \text{ min}^{-1}$ (Fig. 3d). Every other laser, focusing and scanning parameter was identical in these two experiments, showing that the thermal effects are greatly reduced as a result of ablation cooling.

There are numerous applications for soft-tissue ablation^{6,10,14}, particularly in targeting the brain²⁵, where the extreme precision afforded by a laser is of paramount importance. For this reason, we compared the effectiveness of ablation cooling in selective tissue removal from freshly harvested whole rat brains. When the average power is low, heat diffusion from the processing region to the surrounding tissue is low enough that the traditional regime avoids thermal side effects, yielding damage-free ablation¹¹. For higher powers, the ablation-cooled regime demonstrates a clear advantage in the reduction of thermal effects: although low-repetition-rate ablation causes a broad heat-affected zone with damaged neighbouring cells, devascularization and prominent tissue loss (Fig. 3e), there is no major heat damage in the ablation-cooled regime at the same power (600 mW) and pulse energy ($3 \mu\text{J}$) (Fig. 3f). The corresponding ablation rate of $0.75 \text{ mm}^3 \text{ min}^{-1}$ is eight times higher than when using $165 \mu\text{J}$, 180 s pulses, with which a 0.55 mm^3 section of brain tissue was removed in 360 s (ref. 11). With ablation cooling, at a much higher power of 2.7 W (432 MHz intraburst repetition rate, 27 kHz burst repetition rate, $16 \mu\text{J}$ per pulse), virtually thermal-damage-free results are obtained (Fig. 3g) at an ablation rate of $2 \text{ mm}^3 \text{ min}^{-1}$.

Finally, we performed a flap-cutting procedure on a bovine cornea, as this is a realistic indicator for surgical applications⁸. An area several millimetres wide located about 0.4 mm below the surface of the cornea was scanned with the laser and the top layer was then lifted off with a pair of tweezers (Fig. 3h); 24 pulses with $0.8 \mu\text{J}$ of energy per burst were used, which is a reduction by a factor of approximately 15 in pulse fluence compared with previous results⁸. This result and similar experiments on poly(methyl methacrylate) (PMMA) and hydrogels demonstrate that ablation cooling retains several of its benefits even when used for subsurface processing (see Supplementary Information section 15 for a detailed discussion).

We conclude by pointing out three speculative future directions of study: exploration of the far-from-equilibrium thermodynamics of the ablation-cooled regime, whether a suitably sculptured coherent pulse train can coherently enhance nonlinear processes²⁶ and whether similar benefits are possible in proton therapy, because the laser-based generation of bursts of protons seems to be feasible²⁷.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 July 2015; accepted 24 May 2016.

Published online 13 July; corrected online 31 August 2016
(see full-text HTML versions for details).

- Gattass, R. R. & Mazur, E. Femtosecond laser micromachining in transparent materials. *Nat. Photon.* **2**, 219–225 (2008).
- Yang, W., Kazansky, P. G. & Svirko, Y. P. Non-reciprocal ultrafast laser writing. *Nat. Photon.* **2**, 99–104 (2008).
- Steinmeyer, J. D. et al. Construction of a femtosecond laser microsurgery system. *Nat. Protocols* **5**, 395–407 (2010).
- Plech, A., Kotaidis, V., Lorenc, M. & Boneberg, J. Femtosecond laser near-field ablation from gold nanoparticles. *Nat. Phys.* **2**, 44–47 (2006).
- Rousse, A. et al. Non-thermal melting in semiconductors measured at femtosecond resolution. *Nature* **410**, 65–68 (2001).
- Chung, S. H. & Mazur, E. Surgical applications of femtosecond lasers. *J. Biophoton.* **2**, 557–572 (2009).
- Yanik, M. F. et al. Neurosurgery: functional regeneration after laser axotomy. *Nature* **432**, 822 (2004).
- Juhasz, T. et al. Corneal refractive surgery with femtosecond lasers. *IEEE J. Sel. Top. Quant. Electron.* **5**, 902–910 (1999).
- Serbin, J., Bauer, T., Fallnich, C., Kasenbacher, A. & Arnold, W. H. Femtosecond lasers as novel tool in dental surgery. *Appl. Surf. Sci.* **197–198**, 737–740 (2002).
- Hoy, C. L. et al. Clinical ultrafast laser surgery: recent advances and future directions. *IEEE J. Sel. Top. Quant. Electron.* **20**, 242–255 (2014).
- Loesel, F. H. et al. Non-thermal ablation of neural tissue with femtosecond laser pulses. *Appl. Phys. B* **66**, 121–128 (1998).
- Chichkov, B. N., Momma, C., Nolte, S., Alvensleben, F. & Tünnermann, A. Femtosecond, picosecond and nanosecond laser ablation of solids. *Appl. Phys. A* **63**, 109–115 (1996).
- Bauer, F., Michalowski, A., Kiedrowski, T. & Nolte, S. Heat accumulation in ultra-short pulsed scanning laser ablation of metals. *Opt. Express* **23**, 1035–1039 (2015).
- Vogel, A., Noack, J., Hüttman, G. & Paltauf, G. Mechanisms of femtosecond laser nanosurgery of cells and tissues. *Appl. Phys. B* **81**, 1015–1047 (2005).
- Sutton, G. P. & Biblarz, O. *Rocket Propulsion Elements* Ch. 14 (Wiley, 2011).
- Cho, Y. I., Hartnett, J. P. & Rohsenow, W. M. *Handbook of Heat Transfer* 6.21 (McGraw-Hill, 1998).
- Hashida, M. et al. Ablation threshold dependence on pulse duration for copper. *Appl. Surf. Sci.* **197–198**, 862–867 (2002).
- Ancona, A. et al. High speed laser drilling of metals using a high repetition rate, high average power ultrafast fiber CPA system. *Opt. Express* **16**, 8958–8968 (2008).
- Kalaycıoğlu, H., Eken, K. & İlday, F. O. Fiber amplification of pulse bursts up to $20 \mu\text{J}$ pulse energy at 1 kHz repetition rate. *Opt. Lett.* **36**, 3383–3385 (2011).
- Kalaycıoğlu, H. et al. 1 mJ pulse bursts from a Yb-doped fiber amplifier. *Opt. Lett.* **37**, 2586–2588 (2012).
- Kalaycıoğlu, H., Akcaalan, O., Yavas, S., Eldeniz, Y. B. & İlday, F. Ö. Burst-mode Yb-doped fiber amplifier system optimized for low-repetition-rate operation. *J. Opt. Soc. Am. B* **32**, 900–906 (2015).
- Lapczynska, M., Chen, K. P., Herman, P. R., Tan, H. W. & Marjoribanks, R. S. Ultra high repetition rate (133 MHz) laser ablation of aluminum with 1.2-ps pulses. *Appl. Phys. A* **69**, S883–S886 (1999).
- Hu, W., Shin, Y. C. & King, G. Modeling of multi-burst mode pico-second laser ablation for improved material removal rate. *Appl. Phys. A* **98**, 407–415 (2010).
- Marjoribanks, R. S. et al. Ablation and thermal effects in treatment of hard and soft materials and biotissues using ultrafast-laser pulse-train bursts. *Photon. Lasers Med.* **1**, 155–169 (2012).
- Tsai, P. S. et al. All-optical histology using ultrashort laser pulses. *Neuron* **39**, 27–41 (2003).
- Meshulach, D. & Silberberg, Y. Coherent quantum control of two-photon transitions by a femtosecond laser pulse. *Nature* **396**, 239–242 (1998).
- Schwoerer, H. et al. Laser-plasma acceleration of quasi-monoenergetic protons from microstructured targets. *Nature* **439**, 445–448 (2006).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported partially by the European Research Council (ERC) Consolidator Grant ERC-617521 NLL, the European Union (EU) FP7 CROSS TRAP and TÜBITAK under projects 112T980, 112T944 and TEYDEB-3110216. C.K. acknowledges funding from TÜBITAK - BİDEB 2211. We thank Y. Aykaç and V. Aykaç for dental experiments, T. Dalkara, M. Yemişçi Özkan, K. Kılıç for brain tissue experiments, G. Aykut for animal care and brain slicing, I. Mirza, K. Yavuz, G. Makey and M. Karatok for data acquisition and analyses, S. Karahan for histology analyses, A. Büyüksungur and BIOMATEN (METU, Ankara, Turkey) for micro-CT analyses, H. Köymen for PZT characterisation and S. İlday, O. Tokel, H. Çelik, O. Algin and E. Atalar for critical reading of the manuscript.

Author Contributions C.K., H.K. and F.Ö.I. designed the research and interpreted the results. H.K., P.E., S.Y., Ö.A., and C.K. developed the laser systems. H.H. and R.H. developed a high-repetition-rate fibre oscillator. C.K., D.K.K. and B.Ö. performed the laser processing experiments. B.Ç. and C.K. developed the numerical models. M.D.A. carried out brain slicing and histological examinations.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.Ö.I. (ilday@bilkent.edu.tr).

Reviewer Information Nature thanks K. Mitra and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

The majority of the experiments were performed with a customized Yb-doped fibre-laser, which is capable of operating in either a burst or uniform mode at a central wavelength of 1,035 nm. This laser and the other lasers used in the experiments are detailed in Supplementary Information section 3. In burst mode, the laser produces a sequence of an adjustable number of pulses (a burst) with a high intraburst repetition rate. The bursts are repeated at a much lower repetition rate (most commonly 1 kHz or 25 kHz). The intraburst repetition rate of this laser was designed to be switchable between 108 MHz, 216 MHz, 432 MHz, 864 MHz, 1,728 MHz and 3,456 MHz. Lower repetition rates of 1 MHz and 27 MHz could be obtained by selectively picking pulses using the acousto-optic modulator that is used to create the bursts. In uniform mode, the laser produced evenly spaced pulses, typically at 1 kHz or 25 kHz. The pulse durations varied between 300 fs and about 1 ps depending on the pulse energy. Effort was made to keep the pulse durations as similar as possible when making direct comparisons between the burst or uniform modes.

A discussion on the estimation of the minimum repetition rate for the ablation-cooled regime is available in Supplementary Information section 1d. The principle criterion is for the repetition rate of the laser to be faster than the rate at which thermal energy diffuses, or is convected in case of fluids, into the surrounding regions. The dimensions of the interaction volume within which the deposited laser energy needs to be contained can be estimated as the size of the region to be ablated by the subsequent pulses, which is in the range of several hundred nanometres. For highly conductive materials, such as Si, Cu or the ceramic coating of the thermocouple in Fig. 1, we estimate $\tau_0 \approx 1$ ns. (Commonly found values for the thermal relaxation times in the scientific literature pertain to linear absorption, which is not valid for ablation by ultrafast pulses. During ultrafast ablation a plasma state is formed, which greatly changes the absorption properties.) The onset of ablation cooling is gradual (see Supplementary Figs 1 and 3) and even a repetition rate that corresponds to the inverse of $10\tau_0$ confers some of the benefits of this regime. Nevertheless, we have used repetition rates that exceed 1 GHz in most of the experiments that contrast the ablation-cooled regime with the traditional regime.

The preferred method for positioning the laser beam on the sample was to use a computer-controlled galvanometric scanner, owing to their high speeds. Alternatively, the sample can be repositioned using motorized stages. The scanning

speed was adjusted such that a single burst was incident at a given spot. The laser spot size was approximately $24\mu\text{m}$ for most of the experiments. To characterize the ablation efficiency, the scanning speed was adjusted so a single pulse (in the traditional regime) or a single burst (in the ablation-cooled regime) was incident at each ablation spot to eliminate the complicated effects of crater formation and shape on the amount of material ablated. For the experiments that aimed to demonstrate a micromachining procedure, such as drilling, cutting a section of Cu, Si or PbZrTiO₃ (PZT) or the removal of a section of dentine, brain tissue or cornea, multiple scans were performed. This often required the readjustment of the focal plane after each layer of material had been ablated. In these experiments, the absolute durations for the completion of the process depend on the scanning and refocusing parameters. To minimize the influence of such factors, all of the parameters pertaining to the scanning procedure were kept constant when making comparisons between the traditional and ablation-cooled regimes.

Experiments that aimed to compare and contrast the ablation-cooled and traditional regimes were performed on nine different target materials: Si (a semiconductor), Cu (a metal), a thermoelectric module, PZT (a ceramic, which loses its piezoelectricity, when heated), PMMA (a transparent dielectric), dentine (a type of hard tissue) and hydrogel, brain and cornea targets (representative of soft tissues). In the case of (semi-)transparent materials, it is important to ensure that the pulse duration and peak intensities are sufficient to initiate nonlinear absorption. It is also essential that ultrafast pulses (< 10 ps) are used to avoid well-known mechanisms of thermal damage during a pulse.

The processed samples were analysed using bright-field optical microscopy, laser scanning microscopy, scanning electron microscopy and (in several cases) *in situ* optical coherence tomography. Histological analyses were performed using haematoxylin and eosin staining and DAPI staining procedures (Supplementary Information section 12). Soft-tissue experiments were done in accordance with the ethical standards of the Bilkent University Ethics Committee, Approval Number 2013/63. PZT, hydrogel and PMMA experiments are described in Supplementary Information sections 7, 14 and 16, respectively. Details on all of the relevant laser and scanning parameters and the target material properties used in each experiment are provided in the respective sections of Supplementary Information.

Code availability. The code used in the simulations is available on request from the corresponding author.

Sea-ice transport driving Southern Ocean salinity and its recent trends

F. Alexander Haumann^{1,2}, Nicolas Gruber^{1,2}, Matthias Münnich¹, Ivy Frenger^{1,3} & Stefan Kern⁴

Recent salinity changes in the Southern Ocean^{1–7} are among the most prominent signals of climate change in the global ocean, yet their underlying causes have not been firmly established^{1,3,4,6}. Here we propose that trends in the northward transport of Antarctic sea ice are a major contributor to these changes. Using satellite observations supplemented by sea-ice reconstructions, we estimate that wind-driven^{8,9} northward freshwater transport by sea ice increased by 20 ± 10 per cent between 1982 and 2008. The strongest and most robust increase occurred in the Pacific sector, coinciding with the largest observed salinity changes^{4,5}. We estimate that the additional freshwater for the entire northern sea-ice edge entails a freshening rate of -0.02 ± 0.01 grams per kilogram per decade in the surface and intermediate waters of the open ocean, similar to the observed freshening^{1–5}. The enhanced rejection of salt near the coast of Antarctica associated with stronger sea-ice export counteracts the freshening of both continental shelf^{2,10,11} and newly formed bottom waters⁶ due to increases in glacial meltwater¹². Although the data sources underlying our results have substantial uncertainties, regional analyses¹³ and independent data from an

atmospheric reanalysis support our conclusions. Our finding that northward sea-ice freshwater transport is also a key determinant of the mean salinity distribution in the Southern Ocean further underpins the importance of the sea-ice-induced freshwater flux. Through its influence on the density structure of the ocean, this process has critical consequences for the global climate by affecting the exchange of heat, carbon and nutrients between the deep ocean and surface waters^{14–17}.

Observations of salinity in the Southern Ocean over the past few decades have revealed a substantial widespread freshening in the surface waters of both coastal^{10,18} and open ocean regions^{2,5}, as well as in the water masses formed from them^{1,3,4,6}. In particular, the Antarctic Intermediate Water (AAIW) and Subantarctic Mode Water (SAMW) freshened at a rate between -0.01 g kg⁻¹ and -0.03 g kg⁻¹ per decade during the second half of the twentieth century^{1,3,4}. In the Pacific and Indian Ocean sectors, continental shelf waters and the Antarctic Bottom Water (AABW) also freshened substantially^{2,6,10}, while in the Atlantic this freshening was smaller^{6,18}. These salinity changes have been attributed to increased surface freshwater fluxes that stem

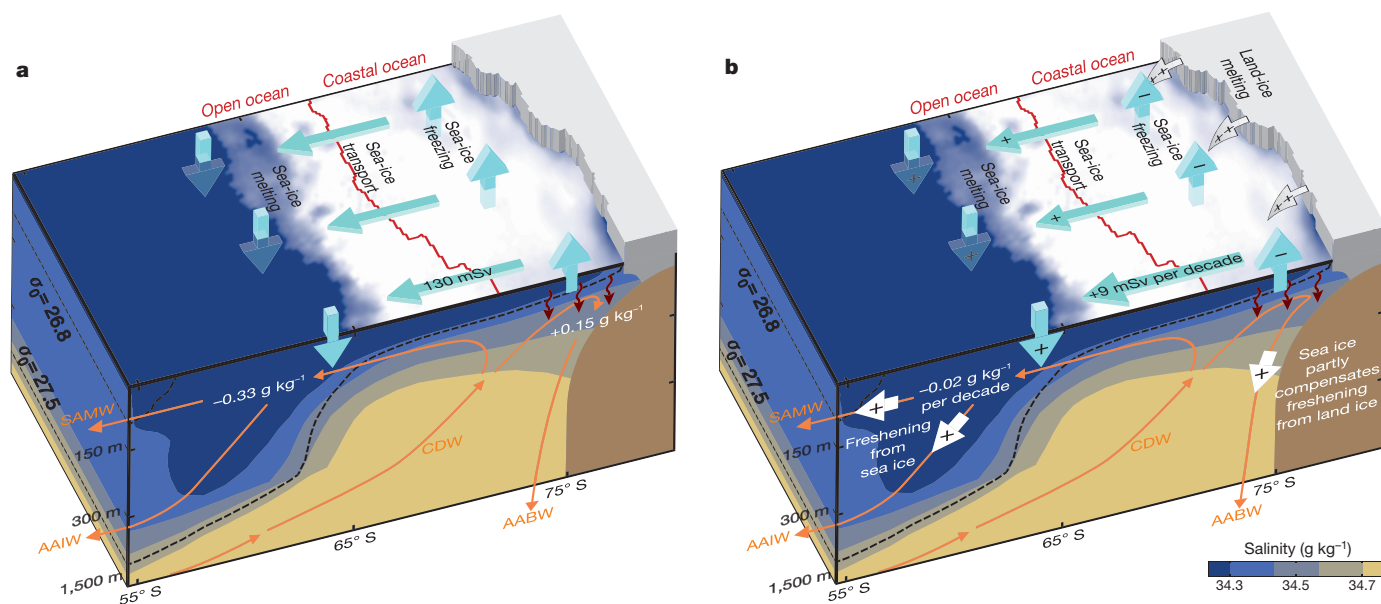


Figure 1 | Effect of northward sea-ice freshwater transport on Southern Ocean salinity. **a, b,** Schematic cross-sections illustrating the effect of northward sea-ice freshwater transport (blue arrows) on mean ocean salinity (**a**) and on the trends over the period 1982 through 2008 (**b**) (see Methods). The red line separates the open and coastal ocean regions. The increasing sea-ice transport freshened the open ocean and, by leaving the salt behind in the coastal region (red curved arrows), compensated for part of the freshening by enhanced glacial meltwater input (grey arrows).

White arrows in **b** indicate the freshening effect from both sea ice and land ice. Positive fluxes are defined downwards or northwards. The orange arrows indicate ocean circulation. The background shows the mean salinity (colour scale) and density (dashed black lines) separating Circumpolar Deep Water (CDW) from Antarctic Intermediate Water (AAIW) and Subantarctic Mode Water (SAMW). AABW, Antarctic Bottom Water.

¹Environmental Physics, Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, Universitätstrasse 16, 8092 Zürich, Switzerland. ²Center for Climate Systems Modeling, ETH Zürich, Universitätstrasse 16, 8092 Zürich, Switzerland. ³Biogeochemical Modelling, GEOMAR Helmholtz Centre for Ocean Research Kiel, Düsternbrooker Weg 20, 24105 Kiel, Germany. ⁴Integrated Climate Data Center (ICDC), Center for Earth System Research and Sustainability, University of Hamburg, Hamburg, Germany.

either from enhanced Antarctic glacial melt^{2,6,10–12} or from increased atmospheric freshwater fluxes, as a result of an excess of precipitation over evaporation^{1,5}. Glacial meltwater¹² is the most likely cause of the freshened coastal waters in the Amundsen and Ross seas^{2,10,11}, but the freshening signal in the AABW, which is formed in this region, is much smaller than expected⁶. In contrast, the recent freshening of the AAIW seems to be much larger than can be explained by the simulated increases in the atmospheric freshwater flux by global climate models in the open Southern Ocean^{1,4}.

Changes in northward sea-ice transport could possibly contribute to the widespread salinity changes in the Southern Ocean⁸. This process acts as a lateral conveyor of freshwater by extracting freshwater from the coastal regions around Antarctica where the sea ice forms and releasing it at the northern edge of the sea ice where the sea ice melts^{19–21} (Fig. 1a). Despite substantial wind-driven changes in sea-ice drift over the past few decades^{8,9}, this contribution has not yet been quantified. Here we suggest that surface freshwater fluxes induced by stronger northward sea-ice transport are a major cause of the observed salinity changes in recent decades; this is corroborated by our finding that the transport process plays a key role in the long-term mean salinity distribution in the Southern Ocean.

Our conclusions are based on basin-scale estimates of annual net sea-ice–ocean freshwater fluxes and the annual northward transport of freshwater by sea ice over the period 1982–2008. Further evidence is provided by our assessment of atmospheric reanalysis data²² and the results from a regional study¹³. We derived the sea-ice-related freshwater fluxes by combining sea-ice concentration, drift and thickness data and by using a mass balance approach to determine the volume divergence and local change in sea ice (Methods). The sea-ice concentration is derived from satellite observations²³ (Extended Data Fig. 1) and its thickness from a combination of satellite data²⁴ and a model-based sea-ice reconstruction that assimilates satellite data²⁵ (Extended Data Fig. 2). The sea-ice volume divergence was computed from satellite-based sea-ice drift vectors²⁶ (Extended Data Figs 3, 4) and sea-ice volume. From the resulting sea-ice volume budget, we estimated the freshwater equivalents of local annual sea-ice–ocean fluxes due to freezing and melting and annual lateral sea-ice transport (Methods).

Uncertainties in these derived freshwater flux products are substantial (Methods). A major challenge arises from the need to combine sea-ice drift estimates from different satellites to estimate the trends. We addressed potential inhomogeneities and biases by vigorous data quality control, implementing several corrections and considering different time periods (Methods). A second challenge is associated with the relatively limited number of observations of sea-ice thickness. These uncertainties plus the observationally constrained range of the other input quantities were incorporated into our error estimates of the final freshwater flux product (Extended Data Tables 1, 2). In the Atlantic sector, uncertainties associated with the mean sea-ice thickness distribution dominate the uncertainty, while in the Pacific sector uncertainties are mostly caused by uncertainties in sea-ice drift.

Our analysis reveals large trends in the meridional sea-ice freshwater transport in the Southern Ocean between 1982 and 2008 (Figs 1b and 2c) that affect the regional sea-ice–ocean freshwater fluxes (Fig. 2d). The annual northward sea-ice freshwater transport of 130 ± 30 mSv ($1 \text{ mSv} = 1,000 \text{ m}^3 \text{ s}^{-1} \approx 31.6 \text{ Gt yr}^{-1}$; Fig. 2a; Extended Data Table 1) from the coastal region to the open ocean strengthened by $+9 \pm 5$ mSv per decade (Extended Data Table 2). Here, the coastal ocean refers to the region between the Antarctic coast and the zero sea-ice–ocean freshwater flux line and the open ocean is the region between the zero sea-ice–ocean freshwater flux line and the sea-ice edge (Fig. 2b). The increased northward transport caused, on average, an additional extraction of freshwater from the coastal ocean of $-40 \pm 20 \text{ mm yr}^{-1}$ per decade and an increased addition to the open ocean region of $+20 \pm 10 \text{ mm yr}^{-1}$ per decade.

The overall intensification occurred primarily in the Pacific sector where we find a vigorous northward freshwater transport trend of

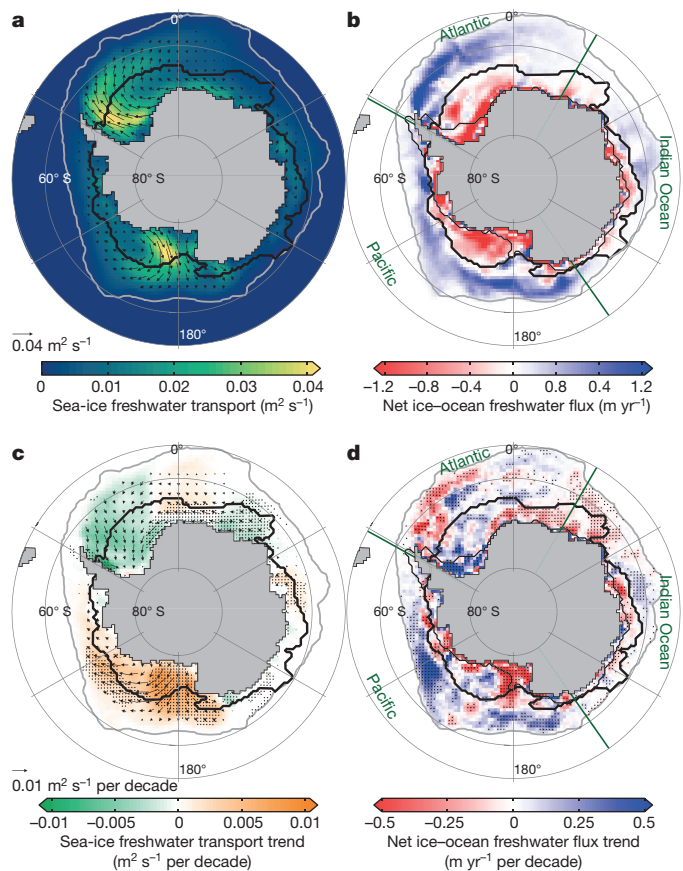


Figure 2 | Mean state and trends of net annual freshwater fluxes associated with sea ice over the period 1982–2008. **a**, Mean sea-ice-induced freshwater transport. **b**, Mean net sea-ice–ocean freshwater flux. **c**, **d**, Linear trends of northward sea-ice freshwater transport (**c**) and net sea-ice–ocean freshwater flux from freezing and melting (**d**). Stippled areas are significant at the 90% confidence level using Student's *t*-test (see Methods). The arrows show the mean (**a**) and trend (**c**) of the annual transport vectors. The thick black lines indicate the zero sea-ice–ocean freshwater flux line that divides the coastal from the open ocean regions, the thin black lines show the continental shelf (1,000 m isobath). The grey lines represent the edge of the sea ice (1% sea-ice concentration) and the green lines show the boundaries of the ocean basins labelled.

$+14 \pm 5$ mSv per decade. The trends in this sector are the most robust (Extended Data Table 3). Over the whole period, this change in the Pacific sector corresponds to an increase of about 30% with respect to the climatological mean in the entire Southern Ocean (Extended Data Table 1). The largest trends occurred locally in the high-latitude Ross Sea (Fig. 2c, d), where our estimated trends agree well with a previous study¹³ (Methods). The increase in the Pacific sector is partly compensated for by small decreases in the Atlantic and Indian ocean sectors. We reach similar conclusions when we consider only the satellite data from 1992 to 2004, that is, the period when they are least affected by potential inhomogeneities (Extended Data Table 3).

The reason for the observed northward sea-ice freshwater transport and its recent trends is the strong southerly winds over the Ross and Weddell seas, which persistently blow cold air from Antarctica over the ocean, pushing sea ice northwards⁹. The winds over the Ross Sea considerably strengthened in recent decades, possibly owing to a combination of natural variability, changes in greenhouse gas concentrations and stratospheric ozone depletion⁹. These changes in the southerly winds induced regional changes in northward sea-ice drift^{8,9}, which are responsible for the sea-ice freshwater transport trends (Methods). This relation between the atmospheric circulation and sea-ice drift changes enabled us to independently estimate the sea-ice

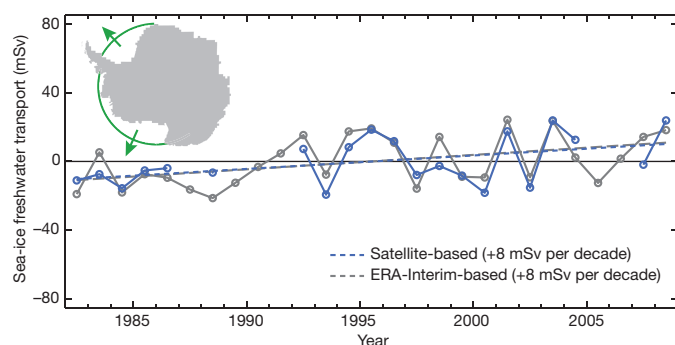


Figure 3 | Time series of annual northward sea-ice freshwater transport anomalies across latitude bands. The underlying sea-ice drift data are based on two independent data sources: the corrected NSIDC satellite data (blue) and zonal sea-level pressure gradients from ERA-Interim data (grey; see Methods). The dashed lines show the respective linear regressions. The map (inset) shows the latitude bands in the Atlantic (69.5° S) and Pacific (71° S) sectors.

drift anomalies using sea-surface pressure gradients along latitude bands from atmospheric reanalysis data²² (Methods). Comparing the resulting northward sea-ice transport anomalies to the satellite-based estimates across the same latitude bands results in a similar overall trend (Fig. 3). Thus, this alternative approach not only corroborates our estimated long-term trend, but also suggests that any remaining inhomogeneities in the sea-ice drift data that are due to changes in the satellite instruments are comparably small after applying multiple corrections (Methods).

To assess how the changing sea-ice–ocean freshwater flux (Fig. 2d) affected the salinity in the Southern Ocean we assumed that the additional freshwater in the open ocean region entered the AAIW and the SAMW formed from upwelling Circumpolar Deep Waters (CDW)^{27,28} (Methods). We find that our freshwater flux trends imply a freshening at a rate of $-0.02 \pm 0.01 \text{ g kg}^{-1}$ per decade in the surface waters that are transported northwards and form the AAIW and SAMW (Fig. 1b). Thus, the sea-ice freshwater flux trend could account for a substantial fraction of the observed long-term freshening in these water masses^{1,3,4}. The strong sea-ice–ocean freshwater flux trends in the Pacific sector (Fig. 2d) spatially coincide with the region of largest observed surface freshening^{2,5} (Extended Data Fig. 7) and can explain also the stronger freshening of the Pacific AAIW compared with that of the Atlantic^{1,4}. A more quantitative attribution of the observed salinity trends to the freshwater transport trends is beyond the scope of our study because the observed freshening trends stem from different time

periods, and have strong regional variations and large uncertainties themselves^{1,3,4}. However, our data show that changes in northward sea-ice freshwater transport induce salinity changes of comparable magnitude to the observed trends.

Our estimates in coastal regions (Fig. 2d) also help to explain the observed salinity changes in the AABW⁶, which is sourced from this region. Additional glacial meltwater from West Antarctica¹² strongly freshened the continental shelf in the Ross and Amundsen seas over recent decades^{2,10,11} (Fig. 1b). However, the observed freshening in Pacific and Indian Ocean AABW was found to be much smaller than expected from this additional glacial meltwater⁶. Our data suggests that the freshening induced by the increasing glacial meltwater is substantially reduced by a salinification from an increased sea-ice to ocean salt flux over the continental shelf in the Pacific sector. This salt flux trend corresponds to a freshwater equivalent of $-10 \pm 3 \text{ mSv}$ per decade, resulting from increasing northward sea-ice export from this region of enhanced sea-ice formation (Fig. 2c, d). In contrast, over the continental shelf in the Atlantic sector our data suggest a decreasing sea-ice to ocean salt flux, corresponding to a freshwater equivalent of $+6 \pm 3 \text{ mSv}$ per decade, which may have contributed to the observed freshening of the newly formed Atlantic AABW⁶ and the north-western continental shelf waters¹⁸.

The large contribution of the trends in sea-ice freshwater transport to recent salinity changes in the Southern Ocean is in line with the dominant role that sea ice plays in the surface freshwater budget in the seasonal sea-ice zone²⁹ and in the global overturning circulation^{19–21,27} in the mean state. The freshwater equivalent of the total Southern Ocean sea-ice melting flux (Fig. 4a) is as large as $460 \pm 100 \text{ mSv}$ (Extended Data Table 1). On an annual basis, the vast majority of this melting flux is supplied by the freezing of seawater of $-410 \pm 110 \text{ mSv}$, with the remaining flux arising from snow-ice formation³⁰ (Methods; Fig. 4b). Most of the sea ice is produced in the coastal region ($-320 \pm 70 \text{ mSv}$), but only about 60% of the sea ice also melts there. The rest, that is, $130 \pm 30 \text{ mSv}$, is exported to the open ocean (Fig. 4c). These mean estimates agree well with an independent parallel study²⁷, which is based on the assimilation of Southern Ocean salinity and temperature observations (Methods).

The process of northward freshwater transport by sea ice effectively removes freshwater from waters that enter the lower oceanic overturning cell, in particular the AABW, and adds it to the upper circulation cell, especially the AAIW (Fig. 1a). Through this process, the salinity difference between these two water masses, and thus the meridional and vertical salinity gradients, increase. In a steady state, the northward sea-ice freshwater transport of $130 \pm 30 \text{ mSv}$ implies a salinity modification of $+0.15 \pm 0.06 \text{ g kg}^{-1}$ and $-0.33 \pm 0.09 \text{ g kg}^{-1}$

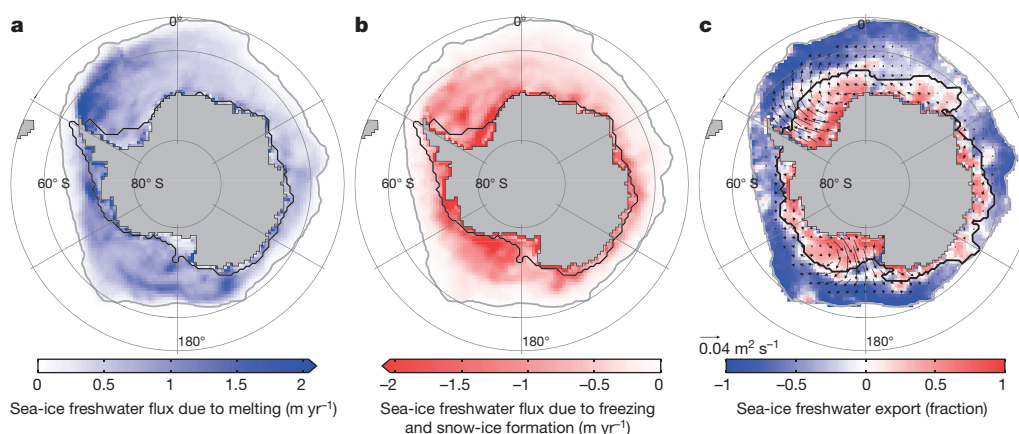


Figure 4 | Mean annual sea-ice-related freshwater fluxes associated with melting, freezing and transport over the period 1982–2008. a, Sea-ice–ocean freshwater flux due to melting. b, Freshwater flux associated with freezing and the formation of snow ice. c, Fraction of freshwater exported

relative to local freezing flux (red) and imported relative to the local melting flux (blue) due to sea-ice induced freshwater transport (arrows). Black and grey lines as in Fig. 2.

in waters that are entering the lower and upper cell, respectively (Methods). The latter suggests that sea-ice freshwater transport accounts for the majority of the salinity difference between the upwelling CDW and the exiting AAIW. We estimated that the salinification from sea ice in waters entering the lower circulation cell is compensated by glacial meltwater and excess precipitation over evaporation in this region in about equal parts, agreeing with the very small salinity difference between the CDW and AABW (Methods).

Because salinity dominates the density structure in polar oceans¹⁴, our findings imply that sea-ice transport is a key factor for the vertical and meridional density gradients in the Southern Ocean and their recent changes (Fig. 1). This interpretation is consistent with the observation that large areas of the upper Southern Ocean not only freshened but also stratified in recent decades⁷. Increased stratification potentially hampers the mixing of deeper, warmer and carbon-rich waters into the surface layer and thus could increase the net uptake of CO₂^{14,16,17}. Consequently, our results suggest that Antarctic sea-ice freshwater transport, through its influence on ocean stratification and the carbon cycle, is more important for changes in global climate^{14,15} than has been appreciated so far. This implication of our findings for the climate system stresses the need to better constrain spatial patterns as well as temporal variations in sea-ice–ocean fluxes by reducing the uncertainties in the observations of drift, thickness and snow cover of Antarctic sea ice.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 November 2015; accepted 8 July 2016.

- Wong, A. P. S., Bindoff, N. L. & Church, J. A. Large-scale freshening of intermediate waters in the Pacific and Indian oceans. *Nature* **400**, 440–443 (1999).
- Jacobs, S. S., Giulivi, C. F. & Mele, P. A. Freshening of the Ross Sea during the late 20th century. *Science* **297**, 386–389 (2002).
- Böning, C. W., Dispert, A., Visbeck, M., Rintoul, S. R. & Schwarzkopf, F. U. The response of the Antarctic Circumpolar Current to recent climate change. *Nat. Geosci.* **1**, 864–869 (2008).
- Helm, K. P., Bindoff, N. L. & Church, J. A. Changes in the global hydrological cycle inferred from ocean salinity. *Geophys. Res. Lett.* **37**, L18701 (2010).
- Durack, P. J., Wijffels, S. E. & Matear, R. J. Ocean salinities reveal strong global water cycle intensification during 1950 to 2000. *Science* **336**, 455–458 (2012).
- Purkey, S. G. & Johnson, G. C. Antarctic Bottom Water warming and freshening: contributions to sea level rise, ocean freshwater budgets, and global heat gain. *J. Clim.* **26**, 6105–6122 (2013).
- de Lavergne, C., Palter, J. B., Galbraith, E. D., Bernardello, R. & Marinov, I. Cessation of deep convection in the open Southern Ocean under anthropogenic climate change. *Nat. Clim. Change* **4**, 278–282 (2014).
- Holland, P. R. & Kwok, R. Wind-driven trends in Antarctic sea-ice drift. *Nat. Geosci.* **5**, 872–875 (2012).
- Haumann, F. A., Notz, D. & Schmidt, H. Anthropogenic influence on recent circulation-driven Antarctic sea ice changes. *Geophys. Res. Lett.* **41**, 8429–8437 (2014).
- Jacobs, S. S. & Giulivi, C. F. Large multidecadal salinity trends near the Pacific–Antarctic continental margin. *J. Clim.* **23**, 4508–4524 (2010).
- Nakayama, Y., Timmermann, R., Rodehacke, C. B., Schröder, M. & Hellmer, H. H. Modeling the spreading of glacial meltwater from the Amundsen and Bellingshausen Seas. *Geophys. Res. Lett.* **41**, 7942–7949 (2014).
- Paolo, F. S., Fricker, H. A. & Padman, L. Volume loss from Antarctic ice shelves is accelerating. *Science* **348**, 327–331 (2015).
- Drucker, R., Martin, S. & Kwok, R. Sea ice production and export from coastal polynyas in the Weddell and Ross Seas. *Geophys. Res. Lett.* **38**, L17502 (2011).
- Sigman, D. M., Hain, M. P. & Haug, G. H. The polar ocean and glacial cycles in atmospheric CO₂ concentration. *Nature* **466**, 47–55 (2010).
- Ferrari, R. *et al.* Antarctic sea ice control on ocean circulation in present and glacial climates. *Proc. Natl Acad. Sci. USA* **111**, 8753–8758 (2014).
- Frölicher, T. L. *et al.* Dominance of the Southern Ocean in anthropogenic carbon and heat uptake in CMIP5 models. *J. Clim.* **28**, 862–886 (2015).
- Landschützer, P. *et al.* The reinvigoration of the Southern Ocean carbon sink. *Science* **349**, 1221–1224 (2015).
- Hellmer, H. H., Huhn, O., Gomis, D. & Timmermann, R. On the freshening of the northwestern Weddell Sea continental shelf. *Ocean Sci.* **7**, 305–316 (2011).
- Saenko, O. A., Schmittner, A. & Weaver, A. J. On the role of wind-driven sea ice motion on ocean ventilation. *J. Phys. Oceanogr.* **32**, 3376–3395 (2002).
- Komuro, Y. & Hasumi, H. Effects of surface freshwater flux induced by sea ice transport on the global thermohaline circulation. *J. Geophys. Res.* **108**, 3047 (2003).
- Kirkman, C. H. & Bitz, C. M. The effect of the sea ice freshwater flux on Southern Ocean temperatures in CCSM3: deep-ocean warming and delayed surface warming. *J. Clim.* **24**, 2224–2237 (2011).
- Dee, D. P. *et al.* The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137**, 553–597 (2011).
- Meier, W. *et al.* NOAA/NSIDC Climate Data Record of Passive Microwave Sea Ice Concentration v. 2, 1980–2009, <http://dx.doi.org/10.7265/N55M63M1> (National Snow and Ice Data Center, accessed 20 June 2013).
- Kurtz, N. T. & Markus, T. Satellite observations of Antarctic sea ice thickness and volume. *J. Geophys. Res.* **117**, C08025 (2012).
- Massonnet, F. *et al.* A model reconstruction of the Antarctic sea ice thickness and volume changes over 1980–2008 using data assimilation. *Ocean Model.* **64**, 67–75 (2013).
- Fowler, C., Emery, W. J. & Tschudi, M. A. Polar Pathfinder Daily 25 km EASE-Grid Sea Ice Motion Vectors v. 2, 1980–2009 (National Snow and Ice Data Center, accessed 14 April 2014).
- Abernathy, R. P. *et al.* Water-mass transformation by sea ice in the upper branch of the Southern Ocean overturning. *Nat. Geosci.* **9**, 596–601 (2016).
- Talley, L. D. Closure of the global overturning circulation through the Indian, Pacific, and Southern Oceans: schematics and transports. *Oceanography* **26**, 80–97 (2013).
- Tamura, T., Ohshima, K. I., Nihashi, S. & Hasumi, H. Estimation of surface heat/salt fluxes associated with sea ice growth/melt in the Southern Ocean. *Sci. Online Lett. Atmos.* **7**, 17–20 (2011).
- Massom, R. A. *et al.* Snow on Antarctic sea ice. *Rev. Geophys.* **39**, 413–445 (2001).

Acknowledgements This work was supported by ETH Research Grant CH2-01 11-1 and by European Union (EU) grant 264879 (CARBOCHANGE). I.F. was supported by C2SM at ETH Zürich and the Swiss National Science Foundation Grant P2EZP2-152133. S.K. was supported by the Center of Excellence for Climate System Analysis and Prediction (CliSAP), University of Hamburg, Germany. F.A.H. and S.K. acknowledge support from the International Space Science Institute (ISSI), Bern, Switzerland, under project #245. We are thankful to F. Massonnet for providing the sea-ice thickness reconstruction and discussion. The ICESat-1 sea-ice thickness data were provided by the NASA Goddard Space Flight Center. The ship-based sea-ice thickness data were provided by the SCAR Antarctic Sea Ice Processes and Climate (ASPeCt) programme. We appreciate the provision of sea-ice concentration and motion data by the National Snow and Ice Data Center, the Integrated Climate Data Center at the University of Hamburg and R. Kwok. We thank T. Frölicher, S. Yang, A. Stössel, M. Frischknecht, L. Papritz, P. Durack, M. van den Broecke, J. Lenaerts, J. van Angelen and M. Meredith for discussion, comments, and ideas.

Author Contributions F.A.H., M.M. and I.F. conceived the study. F.A.H. collated the data and performed the analyses. F.A.H. and N.G. wrote the manuscript. M.M., I.F. and S.K. assisted during the writing process. S.K. assisted in the quality and uncertainty assessment. All authors developed the methods and interpreted the results. N.G. and M.M. supervised this study.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.A.H. (alexander.haumann@usys.ethz.ch).

Reviewer Information Nature thanks K. Ohshima and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Data. The satellite-derived sea-ice concentration is drawn from the Climate Data Record (CDR)²³, which comprises data from the NASA Team algorithm (NTA)³¹ and the Bootstrap algorithm (BA)³², as well as a merged data set. Sea-ice thickness data are taken from a reconstruction with the ocean–sea-ice model NEMO-LIM2 (1980–2009)²⁵, from the laser altimeter ICESat-1 (2003–2008; <http://seaiice.gsfc.nasa.gov>)²⁴, as well as from ship-based observations (ASPeCt; 1980–2005; <http://aspect.antarctica.gov.au>)³³. Satellite-derived sea-ice drift data originates from the National Snow and Ice Data Center (NSIDC)²⁶, is provided in NetCDF-format by the Integrated Climate Data Center (University of Hamburg) and is corrected by drifting buoy data (1989–2005)³⁴. We used an alternative sea-ice drift product for the uncertainty estimation (1992–2003; <http://rkwok.jpl.nasa.gov>; hereafter referred to as Kwok *et al.*)^{35,36}. Additionally, we used daily atmospheric sea-level pressure, surface air temperature and 10 m wind speed values from the ERA-Interim reanalysis (1980–2009, <http://apps.ecmwf.int>)²². We provide a detailed description of the data processing in the corresponding sections below.

Sea-ice concentration. We used all three sea-ice concentration products available from the CDR²³. If any of the grid points in either the merged, NTA or BA products show a sea-ice concentration of 0%, all of the products are set to 0%. We used a first-order conservative remapping method from the Climate Data Operators (CDO)³⁷ to interpolate the sea-ice concentration to the sea-ice drift grid. The BA performs better than the NTA around Antarctica as the NTA underestimates sea-ice concentrations by 10% or more^{23,38} (Extended Data Fig. 1a, b). Therefore, we primarily used the BA product. However, the BA potentially underestimates the concentration of sea ice in the presence of thin ice and leads^{23,38}. Therefore, we used the merged product, which should be more accurate in these regions²³, to estimate the uncertainties. Generally, sea-ice concentration is the best constrained of the three sea-ice variables. Its contribution to the climatological mean flux uncertainty is below 1% (Extended Data Table 1). To obtain the uncertainty in the freshwater flux trends, we also used the NTA because differences in the trends in Antarctic sea-ice area between the BA and NTA have been reported³⁹. Differences between the BA and NTA sea-ice concentration trends range from 10% to 20% relative to the actual trend (Extended Data Fig. 1c, d). The associated uncertainties in the spatially integrated sea-ice freshwater flux trends are about 10% (Extended Data Table 2).

Sea-ice thickness. Sea-ice thickness data spanning our entire analysis period do not exist, mostly owing to challenges in remote sensing of Antarctic sea-ice thickness⁴⁰. We therefore used a sea-ice thickness reconstruction²⁵ from a model that assimilated the observed sea-ice concentration. Through this assimilation, the model constrained air–sea heat fluxes, improving the spatial and temporal variability of the sea-ice thickness. The model did not assimilate sea-ice thickness observations themselves. Sea-ice thickness, as we use it here, is not weighted with sea-ice concentration and does not include the snow layer.

The reconstruction overestimates the sea-ice thickness in the central Weddell and Ross seas and underestimates it in some coastal regions compared to the ICESat-1²⁴ and ASPeCt³³ data sets (Extended Data Fig. 2). To compare the different sea-ice thickness data sets, we interpolated the reconstruction, ICESat-1 and ASPeCt data to the sea-ice drift grid using CDO³⁷ distance-weighted averaging. For our best estimate of the sea-ice freshwater fluxes, we applied a weighted bias correction to the reconstruction using the spatially gridded version of the ICESat-1 data (see the following paragraph). Both the ICESat-1 and ASPeCt data sets are potentially biased low, particularly in areas with thick or deformed sea ice^{33,40–42}, where we found the largest differences between these two data sets and the uncorrected reconstruction. Thus, the thicker sea ice in the Weddell Sea in the uncorrected reconstruction might be realistic, especially when considering alternative ICESat-1 derived estimates for this region^{40,43,44}. To capture the full uncertainty range associated with the mean sea-ice thickness distribution, we used the difference between the uncorrected reconstruction and the ICESat-1 data. Uncertainties in sea-ice thickness dominate the climatological freshwater flux uncertainties in the Atlantic and Indian Ocean sectors, ranging from 10% to 35%, and are also substantial in all other regions and for the overall trends (Extended Data Tables 1, 2).

For the correction of the mean sea-ice thickness distribution, we first calculated relative differences to ICESat-1 whenever data were available. Then, we averaged all of the differences that were within two standard deviations over time. We applied this average relative bias correction map to the data at each time step. To ensure that local extremes were not exaggerated, we used weights. Weights were one for a sea-ice thickness of 1.2 m, that is, the full bias correction was applied, and decreased to zero for sea-ice thicknesses of 0.2 m and 2.2 m, that is, no bias correction was applied. We derived these thresholds empirically to reduce biases with respect to the non-gridded ICESat-1 and ASPeCt data (Extended Data Fig. 2). Trends in the reconstruction remain largely unaffected by the bias correction (comparing Extended Data Fig. 2a and the original trend²⁵).

Local extremes in the sea-ice thickness reconstruction, caused by ridging events, are probably inconsistent with the observed sea-ice drift and would lead to unrealistic short-term variations in our final fluxes. However, when considering the net annual melting and freezing fluxes and averages over large areas these variations cancel out. To reduce the noise in our data set, we filtered extremes with a daily sea-ice thickness anomaly larger than 2 m with respect to the climatological seasonal cycle, representing only 0.1% of all data points. These and other missing grid points (in total 2.6%) were interpolated by averaging the neighbouring grid points. We also calculated our sea-ice freshwater fluxes on the basis of the unfiltered data and included these fluxes in our uncertainty estimate.

Snow-ice formation due to flooding and refreezing^{30,45} is part of the estimated sea-ice thickness. As snow-ice forms partly from the atmospheric freshwater flux and not from the ocean alone, it could lead to an overestimation of the total ocean to sea-ice freshwater flux due to freezing. The amount of snow-ice formation is highly uncertain^{30,45} but lies within the uncertainty of the sea-ice thickness. To account for this process, we reduced the freezing fluxes according to snow-ice formation estimates from the literature³⁰. In the Atlantic, Indian Ocean and Pacific sectors we applied approximate snow-ice formation rates of $8 \pm 8\%$, $15 \pm 15\%$, and $12 \pm 12\%$ of the freezing flux, respectively³⁰. In the entire Southern Ocean, the amount of snow that is transformed to ice would thus amount to about 50 mSv, or about 35% of the suggested atmospheric freshwater flux onto Antarctic sea ice²⁷.

Trends in sea-ice thickness (Extended Data Fig. 2a) are highly uncertain but broadly agree among different modelling studies^{25,46,47}. To show that our results are robust with respect to the less certain trends or short-term variations in sea-ice thickness, we compared our estimated transport trends across the latitude bands (equation (3)) with a sensitivity analysis, where we kept the sea-ice thickness constant. The resulting transport trends across the latitude bands of about -6 mSv per decade in the Atlantic sector and about $+11$ mSv per decade in the Pacific sector are still within our estimated uncertainty (Extended Data Table 2). Most of the sea-ice thickness trends (Extended Data Fig. 2a) occur either north (in the Pacific sector) or south (in the Atlantic sector) of the zero freshwater flux line or latitude bands. Thus, the trend in sea-ice thickness does not considerably affect the northward sea-ice freshwater transport trend. However, the mean sea-ice thickness uncertainty at the zero freshwater flux line is the largest contributor to the overall northward sea-ice freshwater transport trend (Extended Data Table 2).

Sea-ice drift. We used the gridded version of the NSIDC²⁶ sea-ice drift data set. In the Antarctic, it is based on five passive microwave sensors^{48,49} and data from the Advanced Very High Resolution Radiometer (AVHRR)⁵⁰ (Extended Data Fig. 4). Two studies validated this data set with buoy data in the Weddell Sea (1989–2005)³⁴ and around East Antarctica (1985–1997)⁵¹. There is a very high correlation between the buoy and the satellite data on large temporal and spatial scales (that is, monthly and regional) and a strongly reduced agreement on smaller scales (that is, daily and local)^{34,51}. The satellite-derived sea-ice drift underestimates the sea-ice velocity given by the buoys by 34.5%³⁴, that is, faster drift velocities have a larger bias⁵². The bias is smaller for the meridional (26.3%) than for the zonal drift³⁴. We corrected for these low biases by multiplying the drift velocity by the correction factor (1.357) that corresponds to the meridional drift bias³⁴. We argue that the meridional component of the bias is the better estimate in the central sea-ice region, which is the key region for our results. Here, the drift is mainly meridional. The larger biases are observed in the swift, mostly zonal drift along the sea-ice edge that causes the larger zonal biases. The spatial dependence of the bias and our correction imply that larger biases and uncertainties remain in our final product around the sea-ice edge.

We processed this bias-corrected drift data further: first we removed all of the data that were flagged as close to the coast or interpolated over large distances in the product; second, we removed any data with sea-ice concentrations below 50%, closer than 75 km to the coast³⁴, or with a spurious, exact value of zero. Our results are not sensitive to this filtering but it reduces the spatial and temporal noise. After these modifications, about 75% of all of the grid cells covered by sea ice had an associated drift vector.

We compared both the original and the bias-corrected data to a partly independent product by Kwok *et al.*^{35,36}. We interpolated these data onto our grid using CDO³⁷ distance-weighted averaging and applied the same 21-d running mean as for the NSIDC sea-ice drift data. We compared sea-ice drift vectors whenever both data sets were available and sea-ice concentrations were larger than 50%. Extended Data Fig. 3 shows the meridional drift components before and after applying the bias correction factor from the buoy data (Extended Data Fig. 3a and b, respectively). We find that the agreement between the two data sets is much higher after the corrections. Compared with the original NSIDC sea-ice drift data set, the largest improvement occurs in the slope: 1.06 compared with 1.55. Root-mean-square (r.m.s.) differences and the linear correlation coefficient remain identical and the absolute bias is reduced by 0.2 km d^{-1} . The correlation

coefficients between the two data sets are 0.8 for both the zonal and meridional drift components. The spatial patterns of the mean annual sea-ice drift speed (Extended Data Fig. 3c–e) illustrate the improvement in agreement between the two data sets after the application of the bias correction but confirm that considerable differences remain at the sea-ice edge. These differences lead to a relatively high r.m.s. difference in the annual mean sea-ice drift speed in these regions (Extended Data Fig. 3f). However, in the central sea-ice pack—the region that is crucial for our results—the r.m.s. differences are much smaller.

Our bias-corrected sea-ice drift speeds are typically slightly lower (by about 9–19%) than those by Kwok *et al.* but considerably higher than in the uncorrected NSIDC data (about 26%, see above). We used these differences between the data sets to estimate the uncertainties induced by sea-ice drift on the sea-ice freshwater transport (Δu ; Extended Data Tables 1, 2). First, we recomputed all of the fluxes by correcting the original NSIDC data with correction factors derived from the Kwok *et al.* data (1.82 or 45% for the zonal drift, and 1.55 or 35% for the meridional drift) instead of the buoy-derived correction factor. In this way, we also accounted for an uncertainty in the drift direction. Then we averaged the deviations between our best estimate and the estimate based on Kwok *et al.*^{35,36} with those between our best estimate and using the uncorrected and unfiltered NSIDC data. Uncertainties from sea-ice drift in the freshwater fluxes are about 20%. They contribute considerably to the final freshwater flux uncertainty and our trend uncertainties in all regions. **Sea-ice–ocean freshwater flux.** We estimated annual net sea-ice–ocean freshwater fluxes over the period 1982–2008 by calculating the local sea-ice volume change and divergence^{8,53}. From this we derived the local freshwater fluxes F ($\text{m}^3 \text{s}^{-1}$) from the sea ice to the ocean due to freezing and melting on a daily basis through a mass balance:

$$F = -C_{\text{fw}} \left(\frac{\partial(Ach)}{\partial t} + \nabla \cdot (Achu) \right) \quad (1)$$

where the four variables c , h , u and A denote the sea-ice concentration, thickness, drift velocity and grid-cell area, respectively. The factor C_{fw} converts the sea-ice volume flux to a freshwater equivalent⁵⁴:

$$C_{\text{fw}} = \frac{\rho_{\text{ice}}(1 - s_{\text{ice}}/s_{\text{sw}})}{\rho_{\text{fw}}} \quad (2)$$

Here, ρ_{ice} , s_{ice} , s_{sw} and ρ_{fw} are the sea-ice density (925 kg m^{-3})⁵⁵, the sea-ice salinity (6 g kg^{-1})⁵⁶, the reference seawater salinity (34.7 g kg^{-1})²⁸ and the freshwater density ($1,000 \text{ kg m}^{-3}$), respectively.

The annual sea-ice freshwater fluxes were computed from the daily fluxes from March to February of the next year (that is, March 1982 to February 2009), which correspond to the annual freezing and melting cycle of sea ice in the Southern Ocean⁵³. Remaining imbalances between, for example, the open and coastal ocean of the Atlantic sector (Extended Data Tables 1, 2) are due to multiyear sea ice in the coastal region. We performed all of the calculations on the grid of the sea-ice drift data²⁶ and averaged all data products over 3×3 grid boxes, resulting in a nominal resolution of 75 km. To obtain the zero freshwater flux contour line, we averaged the climatological fluxes over 9×9 grid boxes. To estimate the melting and freezing fluxes, we separately summed up the positive and negative daily fluxes over a year (Fig. 4a, b). As temporal fluctuations accumulate when only adding positive or negative values, noise can lead to an overestimation of these fluxes. Each of the sea-ice variables (c , h and u) were therefore low-pass filtered using a 21-d running mean.

Sea-ice freshwater transport. The total northward sea-ice volume transport (in $\text{m}^3 \text{s}^{-1}$) between the coastal and open ocean regions equals the spatial integral of the divergence term in equation (1) in either of the two regions (by Gauss's theorem). We chose the open ocean region because there is considerable zonal exchange between the Indian Ocean and Atlantic sectors (Fig. 2a) in the coastal region, influencing the sector-based estimates. In the open ocean, this effect is negligible. We used this approach for the reported transport estimates (Extended Data Tables 1–3 and Extended Data Fig. 5a–c).

To demonstrate that our main findings are robust on the basin scale, and not influenced by small-scale noise and local uncertainties, we also calculated the northward sea-ice freshwater transport across latitude bands at 69.5°S in the Atlantic sector and 71°S in the Pacific sector (Fig. 3). To this end, we averaged c_n , h_n and meridional drift (v_n) in 1° longitude segments (n) along these latitudes and calculated the local freshwater transport T_n ($\text{m}^3 \text{s}^{-1}$):

$$T_n = C_{\text{fw}} c_n h_n v_n \Delta l_n \quad (3)$$

where Δl_n denotes the length of sectors n along the latitude bands. The combined annual northward freshwater transport of both sectors is $100 \pm 30 \text{ mSv}$ with an increase of $8 \pm 5 \text{ mSv per decade}$ over the period 1982–2008 (Extended

Data Fig. 5d and Fig. 3). This compares well with the mean ($120 \pm 30 \text{ mSv}$) and trend ($9 \pm 5 \text{ mSv per decade}$) of our spatially integrated sea-ice–ocean fluxes in the Pacific and Atlantic (Extended Data Fig. 5b, c).

We calculated the spatial pattern of the sea-ice freshwater transport f ($\text{m}^2 \text{s}^{-1}$) as displayed in Fig. 2a, c, according to:

$$f = C_{\text{fw}} chu \quad (4)$$

Time-series homogenization. Our analysis and earlier studies^{9,57} revealed major temporal inhomogeneities in the NSIDC sea-ice drift data set at the transitions between satellite sensors (Extended Data Fig. 4). We argue that these temporal inhomogeneities are linked to the unavailability of the 85 GHz and 91 GHz channels and sparser data coverage in the earlier years. The drift speed before 1982 seems to be underestimated, which is to some extent mitigated by AVHRR data thereafter. From 1982 to 1986, the drift speed is consistent but has a low bias. The drift ramps up in 1987, when the 85 GHz channels became available, and decreases again between 1989 and 1991, when these channels degraded⁵⁸. A final sudden decrease occurs from 2005 to 2006 when 85 GHz data were not used. We used wind speed data over the sea ice from ERA-Interim²² as an independent data source and scaled it to the sea-ice drift velocity for comparison (Extended Data Figs 4b). The scaling factor stems from the consistent years in the period 1988–2008 and varies in space and with the season^{59,60}. This analysis supports our argument that the sea-ice drift speed is underestimated when the higher resolution 85/91 GHz channels were not available. We note that the meridional drift seems less sensitive to these inhomogeneities than the total drift, which might be related to higher data availability in the central sea-ice pack and is consistent with the lower biases found in the meridional sea-ice drift.

Spurious increases in the sea-ice velocity would affect our estimated trends if they were not taken into account (Extended Data Figs 5, 6). Thus, we corrected the annual divergence (equation (1)) and lateral transport (equations (3), (4)) for the sensor-related temporal inconsistencies as follows. We excluded the inconsistent years (1980, 1981, 1987, 1989–1991, 2005 and 2006) from the analysis. To homogenize the years 1982–1986 with the years 1988–2008, that is, to remove the spurious trend in 1987, we first calculated linear regression lines before and after 1987 at each grid point. Then we added the differences between the end (1986) and start (1988) points of the regression lines to all years before 1987, that is, assuming a zero change in 1987. Fitting regressions before and after spurious jumps is a common procedure to homogenize climate data^{61,62}. Here, we used a linear regression that serves the purpose of computing long-term trends in the time series.

To estimate the sensitivity of the trends in northwards sea-ice freshwater transport to the uncertainties associated with the offset correction before 1987 (shown in orange and green in Extended Data Fig. 5), we performed a Monte Carlo analysis by varying the offset and estimating the resulting trends. We generated 10,000 normally distributed offsets around our best guess (about $19 \pm 5 \text{ mSv}$ for the entire Southern Ocean; Extended Data Table 3). The standard deviation of this distribution was chosen to match the offset uncertainty that arises from the r.m.s. errors of the trends in each of the two time intervals: 1982–1986 and 1988–2008. For each of these generated offsets, we then estimated the trends and their significance (Extended Data Table 3). For both the entire Southern Ocean and the Pacific sector, all of the sampled offsets yield a positive northward sea-ice freshwater transport trend. All trends for the Pacific sector and 92% of those for the entire Southern Ocean are positive and at the same time significant at least at a 90% confidence level using Student's t -test. Thus, our trend results are insensitive to uncertainties in the applied homogenization at the 90% confidence level. The posterior uncertainty shows that the uncertainty associated with the offset has no noticeable effect on the total uncertainty range, that is, is smaller than $\pm 1 \text{ mSv per decade}$.

Uncertainty estimation. The uncertainties of the local (grid-point-based) fluxes and timescales shorter than one year are probably large due to potential inconsistencies between the data sets on such scales and an amplification of the uncertainties by the spatial and temporal differentiations in equation (1). Integrating these terms in space and time greatly reduces these uncertainties (Extended Data Tables 1, 2). We estimated the uncertainties in our product that are associated with the underlying input variables c , h and u by using their observationally constrained ranges from different data sources, including the applied corrections and filtering as described. Additionally, we used an averaging period of 31 d (instead of 21 d) and, for trends only, an estimate without a running-mean filter, to obtain uncertainty estimates associated with temporal noise (Δt). The results confirmed that the annual melting or freezing fluxes, are sensitive to the low-pass filtering, but not the net annual fluxes, as in the latter product the noise is averaged out. The sensitivity of the spatially integrated values to variations of the zero freshwater flux line is estimated by varying the smoothing radius from two to six grid boxes (ΔA). The uncertainty associated with the constant conversion factor (ΔC_{fw} ; equation (2))

is about 5% when using a realistic range of values^{28,55,56}. For the trends only we computed the standard error of the slope from the variance of the residuals around the regression line (Δs_e)⁶³. The total uncertainty for both the climatological mean and the trends was estimated by calculating the r.m.s. of the individual contributions. This analysis shows that in the Atlantic and Indian Ocean sectors both the uncertainties in the climatology and trends (Extended Data Tables 1, 2) are dominated by uncertainties in the sea-ice thickness. In contrast, the uncertainty in the sea-ice drift dominates the uncertainty in the Pacific sector. We tested the significance of the trends with Student's *t*-test, accounting for the fact that only 21 out of 27 years were used and for a lag-1 autocorrelation⁶³. To indicate the significance of the trends at grid-point level (Fig. 2c, d and Extended Data Fig. 6), at which the data uncertainties are unknown, the local r.m.s. of the variance of the residuals was artificially increased by 40%, approximately corresponding to our data uncertainty estimate in Extended Data Table 2. The quality of our data directly at the coastline and around the sea-ice edge is reduced due to the limited quality and quantity of the underlying observations in these regions.

Sea-ice freshwater flux evaluation. A modelling study²⁷ carried out in parallel to this study calculated freshwater fluxes associated with sea-ice formation, melting and transport in the Southern Ocean State Estimate (SOSE). This model assimilates a large amount of observational data and optimizes the surface fluxes. They estimated an annual sea-ice-ocean freshwater flux due to sea-ice formation of -360 mSv over the entire Southern Ocean, which is within our estimated range of -410 ± 110 mSv. Moreover, they estimated that the combined annual sea-ice-ocean freshwater flux due to sea-ice and snow melting is about 500 mSv. Thus, in their estimate a total of 140 mSv of snow accumulated on the sea ice. Our estimates partly include snow accumulation on sea ice, because part of the sea-ice thickness results from snow-ice formation, which we estimated to be about -50 mSv (section on sea-ice thickness). However, the snow layer on top of the sea ice is not included in our estimate of the freshwater flux due to sea-ice melting of 460 ± 100 mSv. In that study²⁷, the authors estimate that the lateral sea-ice freshwater transport from the density class of the CDW to the AAIW and the SAMW amounts to 200 mSv in the period between 2005 and 2010. Their estimate slightly differs from our estimated transport from the coastal to the open ocean, which ranges between about 140 mSv and 160 mSv in 2007 and 2008 (Extended Data Fig. 5). The reasons might be the slightly different regions and that their estimate also includes the transport of the snow layer on top of the sea ice.

Given the reduced confidence in the local fluxes (for example, sea-ice production in coastal polynyas), it is reassuring that our data agree within our estimated range of uncertainty with previous estimates of mean fluxes for some larger coastal polynya regions^{64,65}. Our confidence is higher for fluxes integrated over larger regions, such as the high-latitude Ross and Weddell seas (Extended Data Fig. 5e). Here our estimates are in close agreement with previous studies.

In the Ross Sea, we estimated that the northward transport from the coastal region across a flux gate between Land Bay and Cape Adare³⁶ (the turquoise area in Extended Data Fig. 5e) is 23 ± 5 mSv, increasing by about 30% (or $+7 \pm 4$ mSv) per decade in the period 1992–2008. On the basis of the same passive microwave data, but using a different algorithm for retrieving the sea-ice motion data, two studies^{36,66} found a mean sea-ice area flux across this flux gate of about $1,000,000 \text{ km}^2$ between March and November in the periods 1992–2003 (ref. 36) and 1992–2008 (ref. 66), respectively. Using an approximated mean sea-ice thickness (0.6 m)^{13,66} and the conversion factor (equation (2)), this corresponds to a mean northward freshwater transport of about 19 mSv. In close agreement with our estimate, these studies found an increase of 30% per decade (about $+6$ mSv per decade). Another study¹³, using sea-ice motion from the Advanced Microwave Scanning Radiometer-EOS (AMSR-E), estimated that the mean sea-ice area flux between April and October (2003–2008) across the same flux gate is about $9.3 \times 10^5 \text{ km}^2$ corresponding to a freshwater transport of about 23 mSv. Using the same data, but an alternative approach⁶⁷, they found that the total sea-ice production in all of the Ross Sea polynyas together was about 737 km^3 between April and October (2003–2008), corresponding to a sea-ice-ocean freshwater flux of -31 mSv. This estimate is similar to the total production of about -36 ± 7 mSv south of the flux gate in our data set, because most of the sea-ice production in this region occurs in the polynyas¹³. Using passive microwave data, the same study¹³ found an increase of the production in the Ross Sea polynyas of 28% per decade between 1992 and 2008. A modelling study⁶⁸ found a net annual sea-ice-ocean freshwater flux due to melting and freezing of -27 mSv on the continental shelf in the Ross Sea, which is in agreement with our estimate of -23 ± 5 mSv. They also found a long-term (unquantified, see figure 9b in ref. 68) decrease in the net annual sea-ice-ocean freshwater flux over the Ross Sea continental shelf in the period 1963–2000, which is qualitatively in line with our results.

In the Weddell Sea, the northward sea-ice area flux across a flux gate close to the $1,000 \text{ m}$ isobath (blue area in Extended Data Fig. 5e) has been found to be $5.2 \times 10^5 \text{ km}^2$ on the basis of AMSR-E data between April and October

(2003–2008)¹³. Using an approximated mean sea-ice thickness (0.75 m)¹³ and the conversion factor (2), this corresponds to a mean northward freshwater transport of about 16 mSv. This agrees well with our estimate of an annual northward transport of 16 ± 4 mSv for the same years and the same region. Similar to the Ross Sea, production in the major polynyas of the Weddell Sea was estimated¹³. However, in the Weddell Sea, a large fraction of the sea-ice transported across the flux gate is not produced in the coastal polynyas¹³; thus we cannot directly compare our large-scale estimate to the sea-ice production in the polynyas. In the same study¹³, based on passive microwave data, they found a small, but insignificant long-term decrease in the sea-ice production in the Weddell Sea polynyas between 1992 and 2008, which is qualitatively consistent with our findings in the Atlantic sector. For a much larger area in the Weddell Sea, a modelling study⁶⁹ estimated an annual northward sea-ice freshwater transport of about 34 mSv and another observational study⁷⁰, mostly based on moorings and wind speed, estimated that this flux is as large as about 38 ± 15 mSv. These estimates agree well with our finding of an annual northward freshwater transport of 41 ± 18 mSv across the 69.5°S latitude band, which is approximately their considered transect.

Sea-ice freshwater transport based on ERA-Interim data. To support our findings, we quantified the changes in sea-ice motion that are induced by changes in geostrophic winds^{59,60,70,71} from daily ERA-Interim²² sea-level pressure and surface air temperature data. We averaged the data over 1° longitudinal segments along the previously defined latitude bands (Fig. 3), computed 21-d running means, and smoothed the data spatially over seven longitudinal bins. Then we calculated the sea-level pressure gradients along the latitude bands and used these together with the atmospheric surface density to estimate geostrophic winds normal to the latitude bands^{59,71}. From these, we calculated the sea-ice drift speed using a drift-to-wind-speed ratio of 0.016, derived from drifting buoys in the central Weddell Sea^{59,71}. This parameter is strongly variable in space and time, which is a major uncertainty in the resulting sea-ice drift. Nevertheless, it provides an average estimate for the mostly free drifting sea ice in the central Antarctic sea-ice pack^{59,71}.

The resulting northward sea-ice freshwater transport (equation (3)) is independent in terms of the sea-ice drift but not in terms of the sea-ice concentration and thickness. We used anomalies (at each 1° increment) because the absolute values of the local transport are likely to be biased by the local influences of ocean currents and sea-ice properties. The resulting total annual anomalies of the northward sea-ice freshwater transport agree well in terms of the variability and long-term trend with the transport anomalies based on the satellite sea-ice drift data ($+8$ mSv per decade; Fig. 3). These estimates do not suffer from the temporal inhomogeneities that we identified in the satellite sea-ice drift data (see Methods section ‘Time-series homogenization’).

Sea-ice contribution to ocean salinity. We determined the evolution of ocean salinity s (g kg^{-1}) in response to a given value of F ($\text{m}^3 \text{s}^{-1}$) from a combination of mass and salt balances. The mass balance for a given well-mixed ocean surface box of volume V and density ρ reads:

$$\frac{d\rho V}{dt} = \rho_{\text{in}} Q_{\text{in}} + \rho_{\text{fw}} F - \rho Q_{\text{out}} \quad (5)$$

where Q_{in} and Q_{out} ($\text{m}^3 \text{s}^{-1}$) are the volume fluxes of seawater in and out of the box, ρ_{in} (kg m^{-3}) is the respective density. In a steady state, equation (5) yields:

$$\rho_{\text{in}} Q_{\text{in}} = \rho Q_{\text{out}} - \rho_{\text{fw}} F \quad (6)$$

The corresponding salt balance reads:

$$\rho V \frac{ds}{dt} = \rho_{\text{in}} Q_{\text{in}} s_{\text{in}} - \rho Q_{\text{out}} s \quad (7)$$

We assumed the same constant source water salinity $s_{\text{in}} = s_{\text{sw}}$ and ρ_{fw} as in equation (2) and used a constant reference density ($\rho = 1,027 \text{ kg m}^{-3}$). Moreover, we used the formation rate of the modified water mass as the volume flux of seawater out of the surface box ($Q_{\text{out}} = Q$). Then, substituting equation (6) into equation (7) yields:

$$\rho V \frac{ds}{dt} = (\rho Q - \rho_{\text{fw}} F) s_{\text{sw}} - \rho Q s \quad (8)$$

In a steady state, this results in an equation that describes the modified salinity s as follows:

$$\rho Q s = (\rho Q - \rho_{\text{fw}} F) s_{\text{sw}} \quad (9)$$

Using $s = s_{\text{sw}} + \Delta s$, where Δs is the difference in salinity between the source and modified water masses, equation (9) reduces to:

$$\Delta s = - \frac{\rho_{\text{fw}} s_{\text{sw}} F}{\rho Q} \quad (10)$$

We used net water-mass formation rates (Q) of 29 Sv for formation of the AABW from the CDW and 13 Sv for the formation of the AAIW/SAMW from the CDW²⁸. Figure 1a illustrates the results and shows the zonal mean ocean salinity and density distribution⁷² for comparison.

Assuming that $+130 \pm 30$ mSv of freshwater enter the CDW through northward sea-ice freshwater transport, the salinity modification between the CDW and AAIW/SAMW (using equation (10)) is -0.33 ± 0.09 g kg⁻¹. The uncertainty includes a ± 2 Sv uncertainty in the water-mass formation rate. In observations, the salinity difference between the CDW and the AAIW and SAMW ranges from about -0.3 g kg⁻¹ to -0.5 g kg⁻¹ (ref. 28). Thus, northward freshwater transport by sea-ice could explain the majority of the salinity modification, consistent with very recent findings²⁷ and a mixed-layer salinity budget⁷³.

Similarly, we calculated the contribution of -130 ± 30 mSv of freshwater removed from coastal regions due to northward sea-ice transport to the salinity modification (using equation (10)) between the CDW and AABW, obtaining an increase of $+0.15 \pm 0.06$ g kg⁻¹. The uncertainty includes a ± 7 Sv uncertainty in the AABW formation. However, the observed salinity differences between the CDW and AABW are generally small or even of opposite sign⁷⁴. This is the result of a compensating effect between a sea-ice-driven salinification and a freshening from glacial and atmospheric freshwater. The freshwater fluxes from land ice through basal and iceberg melting are about $+46 \pm 6$ mSv and $+42 \pm 5$ mSv, respectively⁷⁵. Assuming that roughly 60% of the icebergs melt in the coastal regions⁷⁶, a total of about $+70$ mSv are added from the land ice to the coastal ocean, corresponding to a freshening of about -0.08 g kg⁻¹ or a compensation of the sea-ice freshwater flux of about 55% in the AABW. We estimated from the ERA-Interim atmospheric reanalysis data²² that the net atmospheric freshwater flux in the coastal region is about $+80$ mSv, corresponding to a freshening of about -0.09 g kg⁻¹. The resulting net salinity change in coastal waters from sea-ice, atmospheric and land-ice freshwater fluxes is almost zero (-0.02 g kg⁻¹). Such a compensation of the freshwater fluxes in coastal regions was noticed previously^{69,77}. We note that large regional variations of these fluxes have been reported^{75,78}.

To estimate the temporal salinity changes at the surface and in the newly formed AAIW and SAMW, we assumed a constant value of Q and that the freshwater flux and ocean salinity consist of a climatological value plus a time-dependent perturbation ($\bar{F} + F'$ and $\bar{s} + s'$, respectively). Equation (8) then yields:

$$\rho V \frac{ds'}{dt} = \rho Q s_{sw} - \rho_{fw} s_{sw} \bar{F} - \rho Q \bar{s} - \rho_{fw} s_{sw} F' - \rho Q s' \quad (11)$$

As the climatological fluxes are in steady state, the first three terms on the right side in equation (11) cancel according to equation (9), resulting in:

$$\rho V \frac{ds'}{dt} = -\rho_{fw} s_{sw} F' - \rho Q s' \quad (12)$$

We approximated the freshwater flux perturbation ($F' = at$) using our estimated trend a , and rearranged the terms resulting in a first-order linear differential equation:

$$\frac{ds'}{dt} + \frac{Q}{V} s' = -\frac{\rho_{fw} s_{sw} a}{\rho V} t \quad (13)$$

Integration in time yields an expression for the time-dependent evolution of the salinity perturbation:

$$s' = -\frac{\rho_{fw} s_{sw} a}{\rho Q} \left(t - \frac{V}{Q} + \frac{V}{Q} e^{-\frac{Q}{V}t} \right) \quad (14)$$

To obtain an estimate of the salinity trend at a given time t , we substituted equation (14) into equation (13) as follows:

$$\frac{ds'}{dt} = \frac{\rho_{fw} s_{sw} a}{\rho Q} \left(e^{-\frac{Q}{V}t} - 1 \right) \quad (15)$$

The equilibrium response of the system, that is, the long-term trend after several years of perturbation, is:

$$\lim_{t \rightarrow \infty} \frac{ds'}{dt} = -\frac{\rho_{fw} s_{sw} a}{\rho Q} \quad (16)$$

Using our estimated sea-ice freshwater transport trend (a) of $+9 \pm 5$ mSv per decade and an AAIW/SAMW water-mass formation rate as above, we obtained an equilibrium freshening rate of -0.023 ± 0.014 g kg⁻¹ per decade (green in Extended Data Fig. 7b), which is valid for sufficiently large values of Q/V .

Extended Data Fig. 7b (in purple and blue; using equation (14)) shows that if we assumed that the trend started in 1982, there would be a delayed response lowering the mean salinity trend estimate depending on V . We thus tested the sensitivity of the trend to V , which corresponds to the upper 150 m between the zero sea-ice-ocean freshwater flux line and the Subantarctic Front⁷⁹ (Extended Data Fig. 7a), which is the source region of the AAIW. The circumpolar V of about 5×10^6 km³ results in a mean salinity trend (using equation (14)) of -0.014 ± 0.008 g kg⁻¹ per decade between 1982 and 2008 (purple). However, the AAIW formation does not occur in a circumpolar belt but mostly in the south-eastern Pacific and north-western Atlantic, that is, on either side of Drake Passage^{80–84}. Assuming that most of the water is modified in this region and further downstream in the South Pacific^{80,82,84}, we estimated a second, somewhat smaller V of about 2×10^6 km³ (shown in blue). The sea-ice freshwater transport trend into this reference volume is about $+8 \pm 5$ mSv per decade (Figs 2c, d), resulting in a mean salinity trend (using equation (14)) of -0.018 ± 0.010 g kg⁻¹ per decade (blue); because a certain amount of freshwater is transported eastwards out of this sector (blue), the mean trend of the delayed response lies somewhere in between the estimates based on the two different reference volumes (blue and purple).

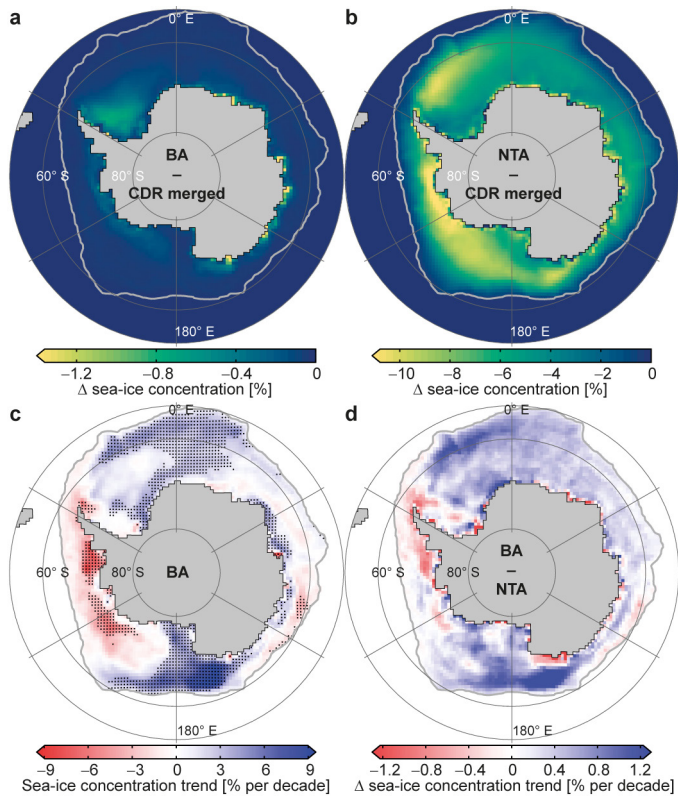
It is unlikely that the trend started exactly in 1982. Thus, the actual salinity response will fall between our estimated delayed response and the equilibrium response. For the range of values above, the deviations in the freshening rate due to effects of a delay and variations in the reference volume are much smaller than the actual magnitude of the trend itself. We thus conclude that the overall mean freshening rate of the newly formed AAIW and the surface waters advected northwards across the Subantarctic Front into the SAMW due to the changes in sea-ice freshwater transport is about -0.02 ± 0.01 g kg⁻¹ per decade (Fig. 1b).

Data deposition. Sea-ice freshwater fluxes leading to the main conclusions are publicly available (<http://dx.doi.org/10.16904/8>). Other presented data are available from the corresponding author upon request.

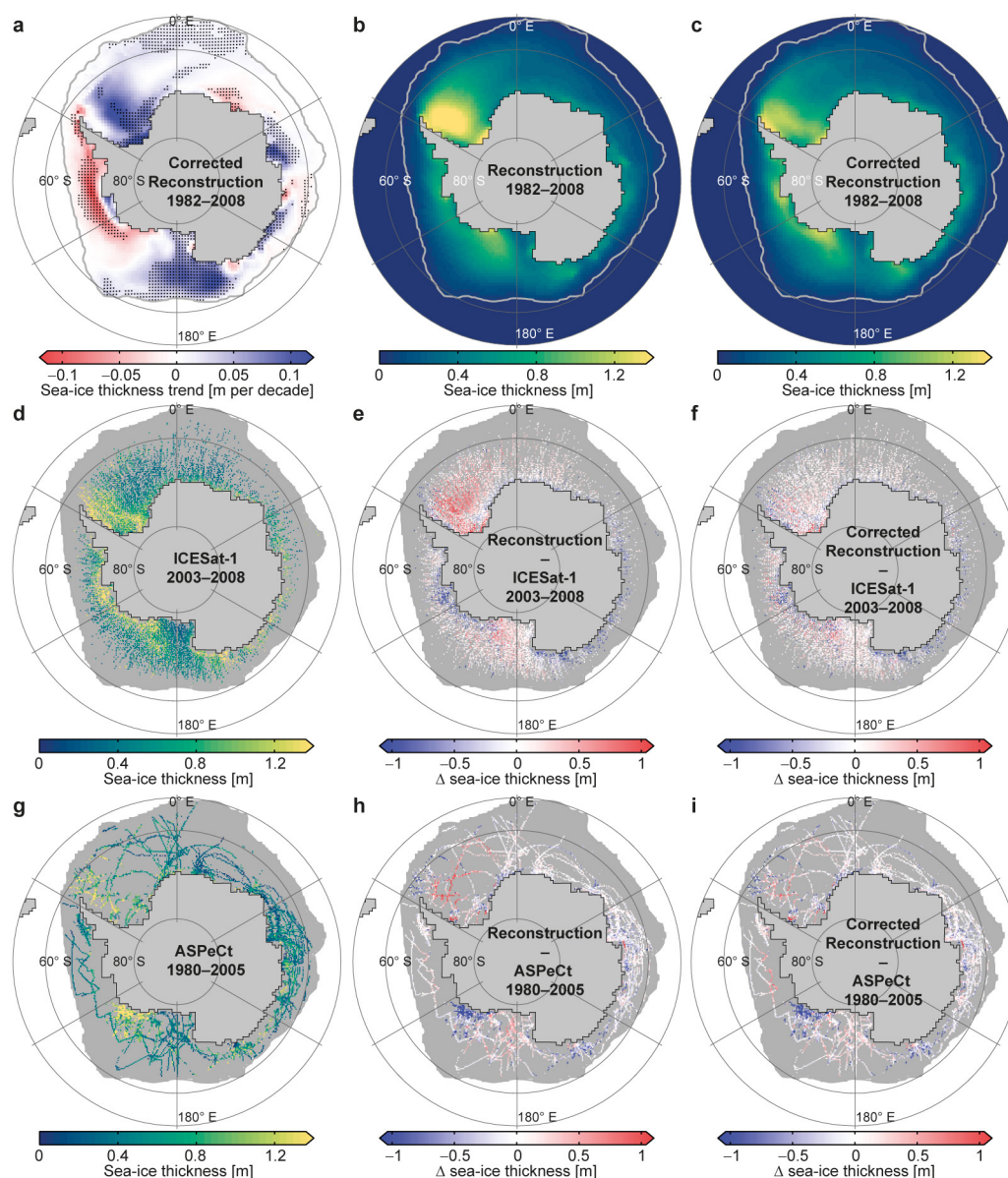
Code availability. Climate Data Operators (CDO; version 1.6.8) used for part of the analysis is publicly available (<http://www.mpimet.mpg.de/cdo>). Other analytical scripts are available upon request from the corresponding author.

31. Cavalieri, D. J. & Parkinson, C. L. Antarctic sea ice variability and trends, 1979–2006. *J. Geophys. Res.* **113**, C07004 (2008).
32. Comiso, J. C. Characteristics of Arctic winter sea ice from satellite multispectral microwave observations. *J. Geophys. Res.* **91**, 975–994 (1986).
33. Worby, A. P. et al. Thickness distribution of Antarctic sea ice. *J. Geophys. Res.* **113**, C05S92 (2008).
34. Schwegmann, S., Haas, C., Fowler, C. & Gerdes, R. A comparison of satellite-derived sea-ice motion with drifting-buoy data in the Weddell Sea, Antarctica. *Ann. Glaciol.* **52**, 103–110 (2011).
35. Kwok, R., Schweiger, A., Rothrock, D. A., Pang, S. & Kottmeier, C. Sea ice motion from satellite passive microwave imagery assessed with ERS SAR and buoy motions. *J. Geophys. Res.* **103**, 8191–8214 (1998).
36. Kwok, R. Ross sea ice motion, area flux, and deformation. *J. Clim.* **18**, 3759–3776 (2005).
37. *Climate Data Operators* v. 1.6.8. (CDO, 2015); <http://www.mpimet.mpg.de/cdo>.
38. Comiso, J. C., Cavalieri, D. J., Parkinson, C. L. & Gloersen, P. Passive microwave algorithms for sea ice concentration: A comparison of two techniques. *Remote Sens. Environ.* **60**, 357–384 (1997).
39. Eisenman, I., Meier, W. N. & Norris, J. R. A spurious jump in the satellite record: has Antarctic sea ice expansion been overestimated? *Cryosphere* **8**, 1289–1296 (2014).
40. Kern, S. & Spreen, G. Uncertainties in Antarctic sea-ice thickness retrieval from ICESat. *Ann. Glaciol.* **56**, 107–119 (2015).
41. Kwok, R. & Maksym, T. Snow depth of the Weddell and Bellingshausen sea ice covers from IceBridge surveys in 2010 and 2011: an examination. *J. Geophys. Res.* **119**, 4141–4167 (2014).
42. Williams, G. et al. Thick and deformed Antarctic sea ice mapped with autonomous underwater vehicles. *Nat. Geosci.* **8**, 61–67 (2015).
43. Yi, D., Zwally, H. J. & Robbins, J. W. ICESat observations of seasonal and interannual variations of sea-ice freeboard and estimated thickness in the Weddell Sea, Antarctica (2003–2009). *Ann. Glaciol.* **52**, 43–51 (2011).
44. Kern, S., Özsoy-Çiçek, B. & Worby, A. Antarctic sea-ice thickness retrieval from ICESat: Inter-comparison of different approaches. *Remote Sens.* **8**, 538 (2016).
45. Maksym, T. & Markus, T. Antarctic sea ice thickness and snow-to-ice conversion from atmospheric reanalysis and passive microwave snow depth. *J. Geophys. Res.* **113**, C02S12 (2008).
46. Zhang, J. Modeling the impact of wind intensification on Antarctic sea ice volume. *J. Clim.* **27**, 202–214 (2014).
47. Holland, P. R. et al. Modeled trends in Antarctic sea ice thickness. *J. Clim.* **27**, 3784–3801 (2014).
48. Emery, W. J., Fowler, C. W. & Maslanik, J. A. in *Oceanographic Applications of Remote Sensing* (eds Ikeda, M. & Dobson, F. W.) 367–379 (CRC Press, 1995).
49. Emery, W. J., Fowler, C. W. & Maslanik, J. A. Satellite-derived maps of Arctic and Antarctic sea ice motion: 1988 to 1994. *Geophys. Res. Lett.* **24**, 897–900 (1997).

50. Maslanik, J. *et al.* AVHRR-based Polar Pathfinder products for modeling applications. *Ann. Glaciol.* **25**, 388–392 (1997).
51. Heil, P., Fowler, C. W., Maslanik, J. A., Emery, W. J. & Allison, I. A comparison of East Antarctic sea-ice motion derived using drifting buoys and remote sensing. *Ann. Glaciol.* **33**, 139–144 (2001).
52. Sumata, H. *et al.* An intercomparison of Arctic ice drift products to deduce uncertainty estimates. *J. Geophys. Res.* **119**, 4887–4921 (2014).
53. Haumann, F. A. *Dynamical Interaction Between Atmosphere and Sea Ice In Antarctica*. MSc thesis, Utrecht University (2011).
54. Ohshima, K. I., Nakanowatari, T., Riser, S., Volkov, Y. & Wakatsuchi, M. Freshening and dense shelf water reduction in the Okhotsk Sea linked with sea ice decline. *Prog. Oceanogr.* **126**, 71–79 (2014).
55. Timco, G. W. & Frederking, R. M. W. A review of sea ice density. *Cold Reg. Sci. Technol.* **24**, 1–6 (1996).
56. Vancoppenolle, M., Fichefet, T. & Goosse, H. Simulating the mass balance and salinity of Arctic and Antarctic sea ice. 2: importance of sea ice salinity variations. *Ocean Model.* **27**, 54–69 (2009).
57. Olason, E. & Notz, D. Drivers of variability in Arctic sea-ice drift speed. *J. Geophys. Res.* **119**, 5755–5775 (2014).
58. Wentz, F. J. *User's Manual: SSM/I Antenna Temperature Tapes Revision 1*. Report No. 120191 (Remote Sensing Systems, 1991).
59. Thorndike, A. S. & Colony, R. Sea ice motion in response to geostrophic winds. *J. Geophys. Res.* **87**, 5845–5852 (1982).
60. Kimura, N. Sea ice motion in response to surface wind and ocean current in the Southern Ocean. *J. Meteorol. Soc. Jpn* **82**, 1223–1231 (2004).
61. Peterson, T. C. *et al.* Homogeneity adjustments of in situ atmospheric climate data: a review. *Int. J. Climatol.* **18**, 1493–1517 (1998).
62. Aguilar, E., Auer, I., Brunet, M., Peterson, T. C. & Wieringa, J. *Guidelines on Climate Metadata and Homogenization*. Report No. WCDMP-53 (World Meteorological Organization, 2003).
63. Santer, B. D. *et al.* Statistical significance of trends and trend differences in layer-average atmospheric temperature time series. *J. Geophys. Res.* **105**, 7337–7356 (2000).
64. Tamura, T., Ohshima, K. I. & Nishihashi, S. Mapping of sea ice production for Antarctic coastal polynyas. *Geophys. Res. Lett.* **35**, L07606 (2008).
65. Ohshima, K. I. *et al.* Antarctic Bottom Water production by intense sea-ice formation in the Cape Darnley polynya. *Nat. Geosci.* **6**, 235–240 (2013).
66. Comiso, J. C., Kwok, R., Martin, S. & Gordon, A. L. Variability and trends in sea ice extent and ice production in the Ross Sea. *J. Geophys. Res.* **116**, C04021 (2011).
67. Martin, S., Drucker, R. S. & Kwok, R. The areas and ice production of the western and central Ross Sea polynyas, 1992–2002, and their relation to the B-15 and C-19 iceberg events of 2000 and 2002. *J. Mar. Syst.* **68**, 201–214 (2007).
68. Assmann, K. M. & Timmermann, R. Variability of dense water formation in the Ross Sea. *Ocean Dyn.* **55**, 68–87 (2005).
69. Timmermann, R., Beckmann, A. & Hellmer, H. H. The role of sea ice in the fresh-water budget of the Weddell Sea, Antarctica. *Ann. Glaciol.* **33**, 419–424 (2001).
70. Harms, S., Fahrbach, E. & Strass, V. H. Sea ice transports in the Weddell Sea. *J. Geophys. Res.* **106**, 9057–9073 (2001).
71. Kottmeier, C. & Sellmann, L. Atmospheric and oceanic forcing of Weddell Sea ice motion. *J. Geophys. Res.* **101**, 20809–20824 (1996).
72. Ingleby, B. & Huddleston, M. Quality control of ocean temperature and salinity profiles — Historical and real-time data. *J. Mar. Syst.* **65**, 158–175 (2007).
73. Ren, L., Speer, K. & Chassignet, E. P. The mixed layer salinity budget and sea ice in the Southern Ocean. *J. Geophys. Res.* **116**, C08031 (2011).
74. Jacobs, S. S. Bottom water production and its links with the thermohaline circulation. *Antarct. Sci.* **16**, 427–437 (2004).
75. Depoorter, M. A. *et al.* Calving fluxes and basal melt rates of Antarctic ice shelves. *Nature* **502**, 89–92 (2013).
76. Silva, T. A. M., Bigg, G. R. & Nicholls, K. W. Contribution of giant icebergs to the Southern Ocean freshwater flux. *J. Geophys. Res.* **111**, C03004 (2006).
77. Jacobs, S. S., Fairbanks, R. G. & Horibe, Y. In *Oceanology of the Antarctic Continental Shelf* (ed. Jacobs, S. S.) 59–85 (American Geophysical Union, 1985).
78. Meredith, M. P. *et al.* Changes in the freshwater composition of the upper ocean west of the Antarctic Peninsula during the first decade of the 21st century. *Prog. Oceanogr.* **87**, 127–143 (2010).
79. Orsi, A. H., Whitworth, T. & Nowlin, W. D. On the meridional extent and fronts of the Antarctic Circumpolar Current. *Deep. Res. I* **42**, 641–673 (1995).
80. England, M. H., Godfrey, J. S., Hirst, A. C. & Tomczak, M. The mechanism for Antarctic Intermediate Water renewal in a world ocean model. *J. Phys. Oceanogr.* **23**, 1553–1560 (1993).
81. Talley, L. D. In *The South Atlantic: Present and Past Circulation* (eds Wefer, G. *et al.*) 219–238 (Springer, 1996).
82. Iudicone, D., Rodgers, K. B., Schopp, R. & Madec, G. An exchange window for the injection of Antarctic Intermediate Water into the South Pacific. *J. Phys. Oceanogr.* **37**, 31–49 (2007).
83. Sloyan, B. M. & Rintoul, S. R. Circulation, renewal, and modification of Antarctic Mode and Intermediate Water. *J. Phys. Oceanogr.* **31**, 1005–1030 (2001).
84. Hartin, C. A. *et al.* Formation rates of Subantarctic mode water and Antarctic intermediate water within the South Pacific. *Deep. Res. I* **58**, 524–534 (2011).
85. Durack, P. J. & Wijffels, S. E. Fifty-year trends in global ocean salinities and their relationship to broad-scale warming. *J. Clim.* **23**, 4342–4362 (2010).

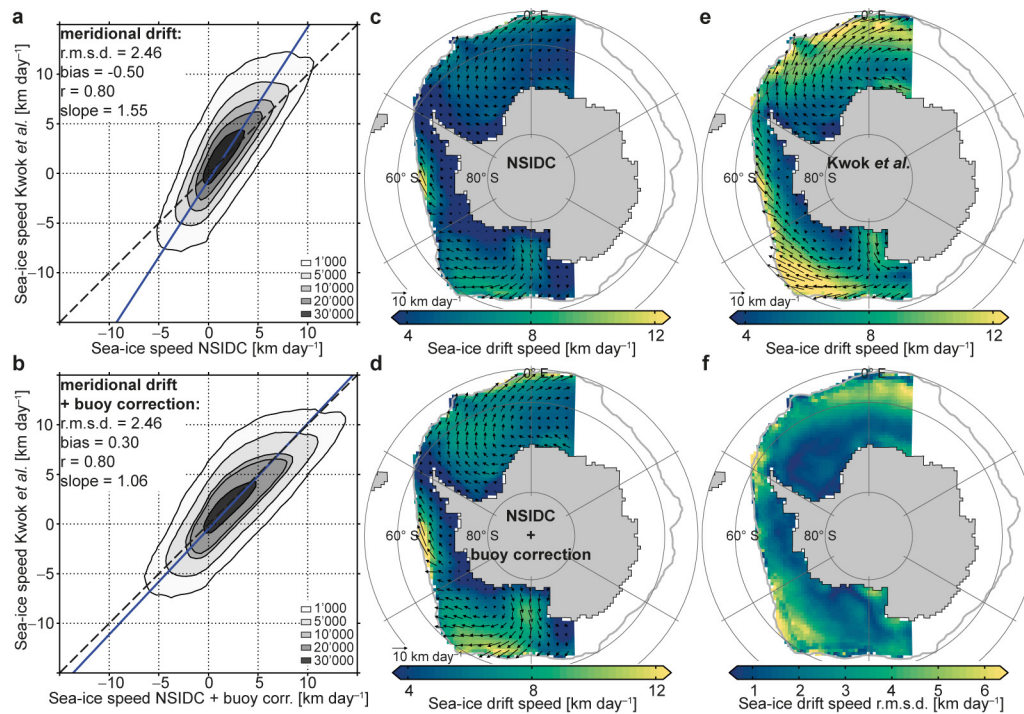


Extended Data Figure 1 | Uncertainties and trends in Antarctic sea-ice concentration over the period 1982–2008. a, BA minus CDR merged data. **b,** NTA minus CDR merged data. **c,** Decadal trends of the BA sea-ice concentration. Stippled trends are statistically significant (at a 90% confidence level or higher using Student's *t*-test). **d,** Decadal trends of the BA minus NTA data. The thick grey line marks the mean sea-ice edge (1% sea-ice concentration). See Methods for details.



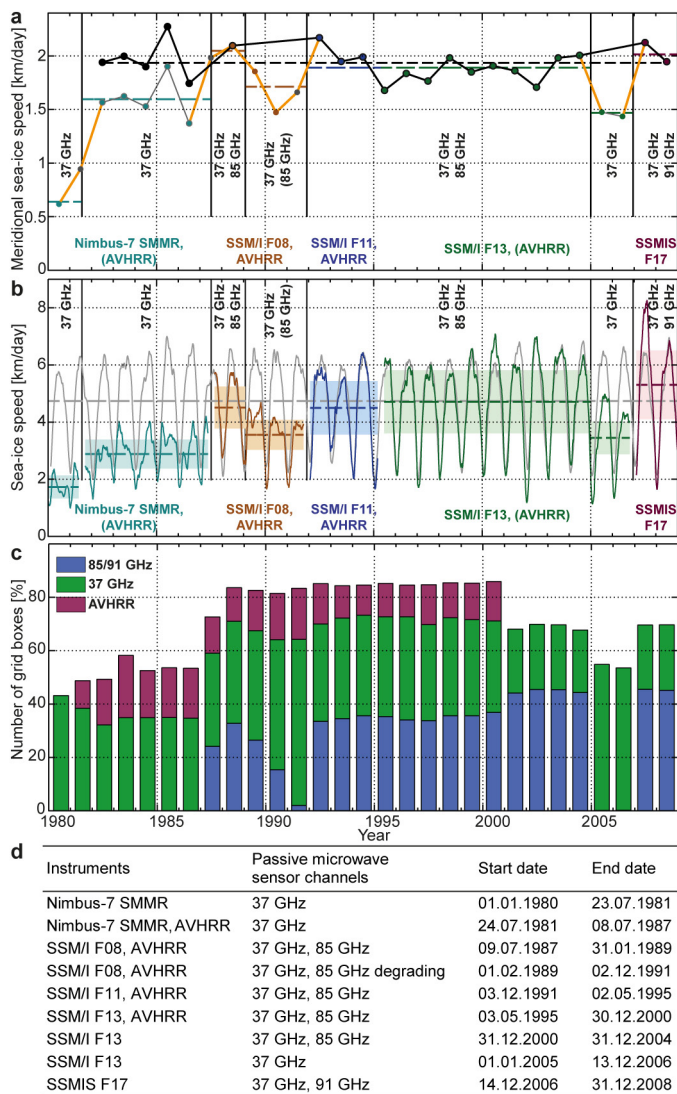
Extended Data Figure 2 | Mean, trend and uncertainty of the Antarctic sea-ice thickness. **a**, Decadal trends of the corrected reconstruction (1982–2008). Stippled trends are statistically significant (at a 90% confidence level or higher using Student's *t*-test). **b**, Mean of the reconstruction (1982–2008). **c**, Mean of the corrected reconstruction (1982–2008). **d**, Mean of the non-gridded ICESat-1 data (2003–2008, 13 campaigns). **e**, Reconstruction minus non-gridded ICESat-1 data (2003–2008). **f**, Corrected reconstruction minus non-gridded

ICESat-1 data (2003–2008). **g**, Mean of the ASPeCt data (1980–2005). **h**, Reconstruction minus ASPeCt data (1980–2005). **i**, Corrected reconstruction minus ASPeCt data (1980–2005). The thick grey line marks the mean sea-ice edge (1% sea-ice concentration). Differences are based on data points when both respective products were available. Data points without data in the sea-ice-covered region are shaded in grey in **d–i**. See Methods for details.

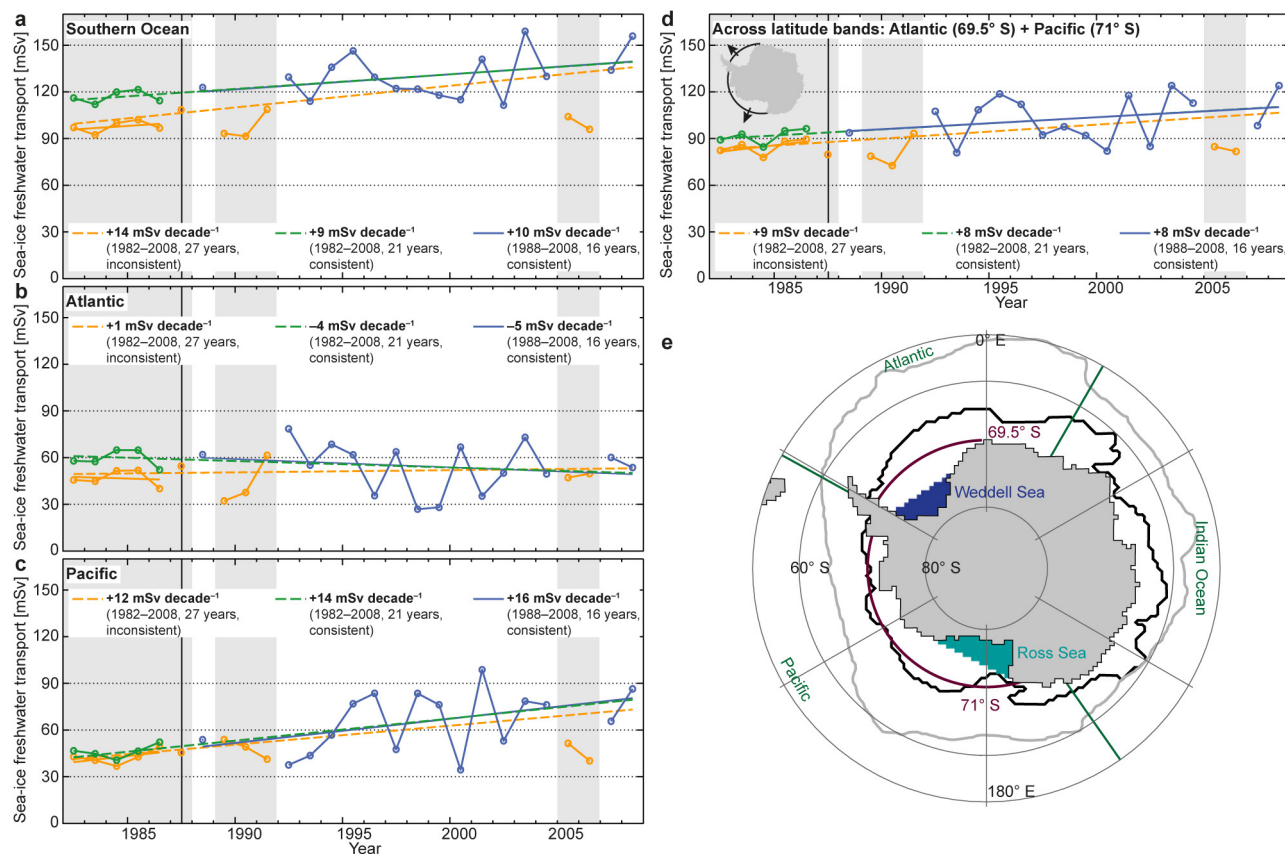


Extended Data Figure 3 | Sea-ice drift speed comparison between the NSIDC and Kwok *et al.* data for the period 1992–2003. **a, b**, Low-pass filtered, 21-d running mean for the original (**a**) and bias-corrected (**b**) daily meridional NSIDC sea-ice drift speed compared with the low-pass filtered daily meridional Kwok *et al.* data. Contours mark the number of grid boxes and the blue line marks the fitted least squares linear regression

line. **c–e**, Mean sea-ice drift speed of the original (**c**) and bias-corrected NSIDC (**d**) and Kwok *et al.* (**e**) sea-ice drift speed. The arrows denote the drift vectors. **f**, R.m.s. differences between the annual mean bias-corrected NSIDC and Kwok *et al.* sea-ice drift speed. The thick grey line in **c–f** marks the mean sea-ice edge (1% sea-ice concentration). Data points were compared when both data sets were available. See Methods for details.

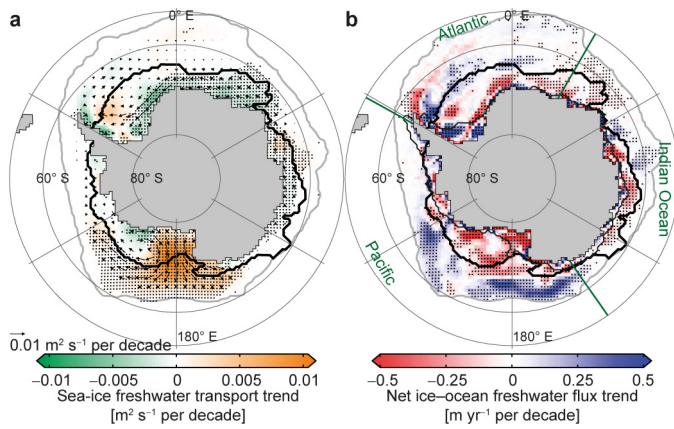


Extended Data Figure 4 | Temporal inhomogeneities in the NSIDC satellite sea-ice drift data. **a**, Annual mean meridional sea-ice drift speed averaged over the entire sea-ice area (sea-ice concentration >50%). The thick orange lines show the spurious trends due to changes in the underlying data. The black lines show the data corrected for inconsistencies and used in this study (1982–2008). **b**, Low-pass filtered (91 d running mean) sea-ice drift speed averaged over the entire sea-ice area (sea-ice concentration >50%). The grey lines show the reduced wind speed from ERA-Interim using a reduction factor from the period 1988–2008. The uncorrected data for each satellite instrument combination are shown in colour (dashed lines show the mean over the respective period). The black vertical lines show the periods of the channels. The coloured text denotes the sensors and the frequency of the microwave radiometer channels used. **c**, The fraction of sea-ice covered grid boxes with at least one drift vector observation in a 21-d window and a 75 km × 75 km grid box using the non-gridded NSIDC drift data. The colours indicate the contribution of each sensor and channel. **d**, Different combinations of instruments and passive microwave sensor channels and the related periods underlying the NSIDC sea-ice drift data. See Methods for details.



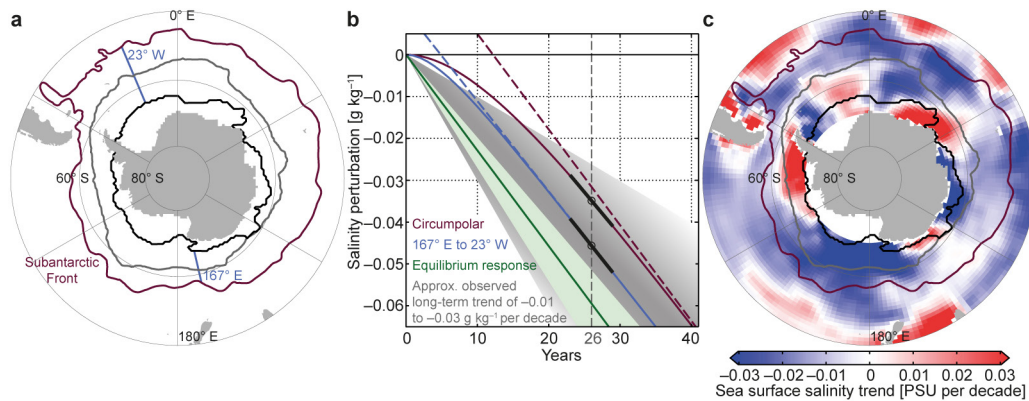
Extended Data Figure 5 | Time series and regions of annual northward sea-ice freshwater transport. **a–c**, Transport from the coastal ocean to the open ocean region in the Southern Ocean (**a**), Atlantic sector (**b**) and Pacific sector (**c**). **d**, Transport across latitude bands in the Atlantic (69.5° S) and Pacific (71° S) sectors. Orange indicates transport estimates if temporal inhomogeneities were not accounted for. Blue shows homogeneous years only. Green represents homogenized time series. Years that have been corrected or removed are shaded in grey. Straight lines show the linear regressions for the periods 1982–2008 (dashed orange and green), 1982–1986 (solid orange) and 1988–2008 (homogeneous

years only; solid blue). See Methods for details. **e**, Regions used for the evaluation of the sea-ice freshwater fluxes. Turquoise shading indicates the area south of the coastal Ross Sea flux gate^{13,36,66}. Dark blue shading highlights the area south of the coastal Weddell Sea flux gate¹³. Purple lines are the 69.5° S latitude band in the Atlantic sector and the 71° S latitude band in the Pacific sector. The black line shows the smoothed mean zero sea-ice-ocean freshwater flux line that divides the coastal and open ocean regions (see Methods). The thick grey line shows the mean sea-ice edge (1% sea-ice concentration) and the green lines mark basin boundaries.



Extended Data Figure 6 | Trends of the net annual freshwater fluxes associated with sea ice over the period 1982–2008 if temporal inhomogeneities in the sea-ice drift data were not considered.

a, b, Linear trends in the meridional sea-ice freshwater transport (**a**) and the net sea-ice–ocean freshwater flux from freezing and melting (**b**). The arrows in **a** denote the trend of the annual transport vectors. Stippled trends are significant at the 90% confidence level using Student's *t*-test (Methods). Thick black lines show the zero sea-ice–ocean freshwater flux line used to divide the coastal from the open ocean regions; the thin black lines mark the continental shelf (1,000 m isobath) the grey lines show the sea-ice edge (1% sea-ice concentration) and the green lines indicate the basin boundaries.



Extended Data Figure 7 | Contribution of sea-ice freshwater flux trends to ocean salinity. **a**, Map showing the regions used for the estimation of salinity changes due to sea-ice freshwater fluxes. The blue lines show the sector important for AAIW formation (167° E to 23° W). The purple line is the Subantarctic Front⁷⁹. The black line indicates the smoothed mean zero freshwater flux line that divides the coastal and open ocean regions. The thick grey line is the mean sea-ice edge (1% sea-ice concentration). **b**, The salinity response to a freshwater flux perturbation using the

long-term equilibrium response (green) and using a delayed response starting in 1982 for a circumpolar reference volume ($5 \times 10^6 \text{ km}^3$; purple) or for the region of most AAIW formation ($2 \times 10^6 \text{ km}^3$; blue). See Methods for details. Dashed lines show the respective asymptotic equilibrium response. The black lines are the respective current trends. The grey shading shows the approximate observed long-term trend in the AAIW^{1,3,4}. **c**, Observed long-term sea-surface salinity trends (data from ref. 85).

Extended Data Table 1 | Mean and uncertainties of the annual sea-ice freshwater fluxes over the period 1982–2008

	Flux [mSv]	Δt [mSv]	ΔA [mSv]	Δc [mSv]	Δh [mSv]	Δu [mSv]	ΔC_{fw} [mSv]
Southern Ocean:							
Transport	+130 ±30	±0	±5	±0	±16	±25	±6
Net open ocean	+130 ±30	±0	±5	±0	±16	±25	±6
Net coastal ocean	-130 ±30	±0	±5	±0	±14	±26	±6
Net continental shelf	-60 ±20	±0	±0	±0	±8	±13	±3
Total melting	+460 ±100	±37	-	±1	±74	±49	±23
Total freezing	-410 ±110	±37	-	±1	±73	±50	±23
Atlantic sector:							
Transport	+60 ±20	±0	±1	±0	±13	±11	±3
Net open ocean	+60 ±20	±0	±1	±0	±13	±11	±3
Net coastal ocean	-50 ±20	±0	±1	±0	±14	±9	±3
Net continental shelf	-20 ±5	±0	±0	±0	±2	±4	±1
Total melting	+180 ±40	±13	-	±0	±25	±21	±9
Total freezing	-160 ±40	±13	-	±0	±25	±19	±9
Indian Ocean sector:							
Transport	+10 ±5	±0	±1	±0	±4	±2	±1
Net open ocean	+10 ±5	±0	±1	±0	±4	±2	±1
Net coastal ocean	-10 ±6	±0	±1	±0	±4	±4	±1
Net continental shelf	-10 ±4	±0	±0	±0	±3	±2	±0
Total melting	+70 ±30	±7	-	±0	±24	±5	±4
Total freezing	-70 ±30	±7	-	±0	±24	±6	±4
Pacific sector:							
Transport	+60 ±20	±0	±2	±0	±9	±12	±3
Net open ocean	+60 ±20	±0	±2	±0	±9	±12	±3
Net coastal ocean	-60 ±20	±0	±2	±0	±9	±13	±3
Net continental shelf	-30 ±9	±0	±0	±0	±6	±6	±2
Total melting	+200 ±50	±17	-	±0	±43	±23	±10
Total freezing	-180 ±60	±17	-	±0	±43	±24	±10

Positive numbers indicate a freshwater flux into the ocean or northward transport (1 mSv = $10^3 \text{ m}^3 \text{ s}^{-1}$). The final uncertainty estimate (95% confidence level) stems from the uncertainties in the filtering of high-frequency temporal noise (Δt), variations of the zero freshwater flux line (ΔA), sea-ice concentration (Δc), sea-ice thickness (Δh), sea-ice drift (Δu) and the freshwater conversion factor (ΔC_{fw}), respectively. See Methods for details. See Fig. 2 for the definition of regions.

Extended Data Table 2 | Decadal trends of the annual sea-ice freshwater fluxes and their uncertainties over the period 1982–2008

	Flux [mSv dec ⁻¹]	Δs_e [mSv dec ⁻¹]	Δt [mSv dec ⁻¹]	ΔA [mSv dec ⁻¹]	Δc [mSv dec ⁻¹]	Δh [mSv dec ⁻¹]	Δu [mSv dec ⁻¹]	ΔC_{fw} [mSv dec ⁻¹]
Southern Ocean:								
Transport	+9 ±5	±3.2	±0.3	±1.1	±0.8	±3.0	±1.9	±0.5
Net open ocean	+10 ±5	±3.5	±0.4	±1.1	±0.8	±3.0	±2.0	±0.5
Net coastal ocean	-10 ±5	±3.5	±0.2	±1.1	±0.7	±3.3	±1.1	±0.5
Net continental shelf	-3 ±2	±1.8	±0.0	±0.0	±0.1	±0.8	±0.1	±0.1
Atlantic sector:								
Transport	-4 ±5	±4.3	±0.1	±0.7	±0.1	±1.4	±0.7	±0.2
Net open ocean	-4 ±5	±4.4	±0.1	±0.7	±0.1	±1.4	±0.7	±0.2
Net coastal ocean	+6 ±6	±5.7	±0.1	±0.7	±0.0	±0.6	±1.8	±0.3
Net continental shelf	+6 ±3	±2.5	±0.0	±0.0	±0.0	±0.6	±1.6	±0.3
Indian Ocean sector:								
Transport	-1 ±1	±1.3	±0.0	±0.2	±0.1	±0.3	±0.2	±0.0
Net open ocean	-1 ±1	±1.3	±0.0	±0.2	±0.1	±0.3	±0.2	±0.0
Net coastal ocean	-3 ±2	±0.9	±0.0	±0.2	±0.1	±1.1	±0.7	±0.1
Net continental shelf	+2 ±1	±0.9	±0.1	±0.0	±0.1	±0.3	±0.4	±0.1
Pacific sector:								
Transport	+14 ±5	±3.4	±0.2	±0.6	±0.7	±1.3	±2.8	±0.7
Net open ocean	+14 ±5	±3.4	±0.3	±0.5	±0.7	±1.2	±2.9	±0.7
Net coastal ocean	-13 ±5	±3.6	±0.2	±0.5	±0.6	±1.9	±2.3	±0.7
Net continental shelf	-10 ±3	±2.6	±0.1	±0.0	±0.2	±1.2	±1.8	±0.5

Positive numbers indicate a freshwater flux trend into the ocean or a northward transport trend (1 mSv per decade = $10^9 \text{ m}^3 \text{ s}^{-1}$ per decade). The final uncertainty estimate (95% confidence level) stems from the standard error of the slope of the regression line (Δs_e), filtering of high-frequency temporal noise (Δt), variations of the zero freshwater flux line (ΔA), sea-ice concentration (Δc), sea-ice thickness (Δh), sea-ice drift (Δu) and the freshwater conversion factor (ΔC_{fw}), respectively. Bold numbers indicate a significance of at least 90% confidence using Student's *t*-test. See Methods for details. See Fig. 2 for the definition of the regions.

Extended Data Table 3 | Sensitivity of the northward sea-ice freshwater transport trend to time periods and homogenization

	Southern Ocean	Atlantic sector	Indian Ocean sector	Pacific sector
1992 – 2004:				
Flux trend [mSv dec ⁻¹]	+4 ±9	-12 ±11	-5 ±3	+21 ±10
1992 – 2008:				
Flux trend [mSv dec ⁻¹]	+11 ±8	-5 ±9	-2 ±2	+17 ±8
1982 – 2004:				
Flux trend [mSv dec ⁻¹]	+8 ±5	-6 ±5	-1 ±1	+15 ±6
1982 – 2008:				
Flux trend [mSv dec ⁻¹]	+9 ±5	-4 ±5	-1 ±1	+14 ±5
1982 – 2008				
Monte Carlo analysis:				
Flux offset before 1987 [mSv]	+19 ±5	+13 ±7	+3 ±2	+4 ±5
Probability for trend of same sign [%]	100	92	78	100
Probability for significant trend of same sign [%]	92	26	9	100
Posterior trend uncertainty [mSv dec ⁻¹]	±5	±6	±2	±5

Positive numbers indicate a northward freshwater transport trend (1 mSv per decade = $10^3 \text{ m}^3 \text{ s}^{-1}$ per decade). Bold numbers indicate a significance of the trend of at least 90% confidence using Student's *t*-test. The Monte Carlo analysis is performed for 10,000 normally distributed sample offsets. Uncertainties (at the 95% confidence level) stem from the standard error of the slope of the regression line and the data uncertainty. See Methods for details. See Fig. 2 for the definition of the regions.

Addition of multiple limiting resources reduces grassland diversity

W. Stanley Harpole^{1,2,3}, Lauren L. Sullivan⁴, Eric M. Lind⁴, Jennifer Firn⁵, Peter B. Adler⁶, Elizabeth T. Borer⁴, Jonathan Chase^{2,3}, Philip A. Fay⁷, Yann Hautier⁸, Helmut Hillebrand⁹, Andrew S. MacDougall¹⁰, Eric W. Seabloom⁴, Ryan Williams¹¹, Jonathan D. Bakker¹², Marc W. Cadotte¹³, Enrique J. Chanton¹⁴, Chengjin Chu¹⁵, Elsa E. Cleland¹⁶, Carla D'Antonio¹⁷, Kendi F. Davies¹⁸, Daniel S. Gruner¹⁹, Nicole Hagenah²⁰, Kevin Kirkman²⁰, Johannes M. H. Knops²¹, Kimberly J. La Pierre²², Rebecca L. McCulley²³, Joslin L. Moore²⁴, John W. Morgan²⁵, Suzanne M. Prober²⁶, Anita C. Risch²⁷, Martin Schuetz²⁷, Carly J. Stevens²⁸ & Peter D. Wragg²⁹

Niche dimensionality provides a general theoretical explanation for biodiversity—more niches, defined by more limiting factors, allow for more ways that species can coexist¹. Because plant species compete for the same set of limiting resources, theory predicts that addition of a limiting resource eliminates potential trade-offs, reducing the number of species that can coexist². Multiple nutrient limitation of plant production is common and therefore fertilization may reduce diversity by reducing the number or dimensionality of belowground limiting factors. At the same time, nutrient addition, by increasing biomass, should ultimately shift competition from belowground nutrients towards a one-dimensional competitive trade-off for light³. Here we show that plant species diversity decreased when a greater number of limiting nutrients were added across 45 grassland sites from a multi-continent experimental network⁴. The number of added nutrients predicted diversity loss, even after controlling for effects of plant biomass, and even where biomass production was not nutrient-limited. We found that elevated resource supply reduced niche dimensionality and diversity and increased both productivity⁵ and compositional turnover. Our results point to the importance of understanding dimensionality in ecological systems that are undergoing diversity loss in response to multiple global change factors.

The search for the mechanisms underlying the coexistence of multiple species was inspired by Darwin's observations of the problem of the 'entangled bank', or how different checks on the growth of individuals underlie the number of species found together⁶. One of the most general theoretical explanations for this problem is that greater dimensionality, or number of non-overlapping ecological niches, allows for the coexistence of a greater number of species^{1,7}. However, plant coexistence challenges this understanding: rather than occupying unique resource niches, plants share and are limited by the same essential resources⁸. The coexistence of plants competing for the same resources therefore requires stoichiometric and physiological trade-off

differences for shared limiting resources². Furthermore, plant resources are spatially separated, with elemental nutrients (for example, nitrogen, phosphorus, potassium) and water acquired belowground and light aboveground. This suggests that two, non-independent resource-based mechanisms could maintain plant diversity: multi-dimensional trade-offs for belowground limiting nutrients, juxtaposed with a one-dimensional trade-off for light aboveground.

Resource competition theory predicts that addition of a limiting resource makes that resource non-limiting, thereby eliminating a competitive trade-off contributing to coexistence². Because some factor must ultimately limit growth, resource additions will lead to a reduction in the number and a shift in the identity of growth-limiting factors. In the case of plants, addition of multiple nutrients should reduce the dimensionality of belowground resource trade-offs, increase biomass production, and ultimately shift the prevailing form of resource competition towards a single, aboveground limiting resource, light^{3,5}. Support for this hypothesis has been demonstrated in four grassland experiments. All of these experiments found plant biomass production was limited by multiple resources, and diversity decreased as a function of the number of belowground resources made non-limiting^{5,9–11}. These results are consistent with the hypothesis that multi-dimensional trade-offs for belowground resources, and light competition mediated by aboveground biomass production, might jointly contribute to maintaining plant diversity in natural communities. Although multiple limitation of primary producer communities is common¹², a recent global study demonstrated substantial site-level variation in the number and identity of co-limiting resources, with around 25% of sites showing no evidence that biomass production was nutrient limited¹³. The question remains whether the dimensionality of nutrient resources might contribute to plant diversity independently of the presumed importance of indirect effects of biomass on diversity.

Here we tested for loss of species diversity in response to multiple nutrient additions⁵ using the Nutrient Network, a globally-distributed,

¹Department of Physiological Diversity, Helmholtz Center for Environmental Research – UFZ, Permoserstrasse 15, Leipzig 04318, Germany. ²German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, Leipzig 04103, Germany. ³Institute of Biology, Martin Luther University Halle-Wittenberg, Am Kirchtor 1, Halle (Saale) 06108, Germany. ⁴Department of Ecology, Evolution, and Behavior, University of Minnesota, St Paul, Minnesota 55108, USA. ⁵School of Earth, Environment and Biological Sciences, Queensland University of Technology, Brisbane, Queensland 4001, Australia. ⁶Department of Wildland Resources and the Ecology Center, Utah State University, Logan, Utah 84322, USA. ⁷USDA-ARS Grassland Soil and Water Research Lab, Temple, Texas 76502, USA. ⁸Ecology and Biodiversity Group, Department of Biology, Utrecht University, Padualaan 8, Utrecht, CH 3584, The Netherlands. ⁹Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, Schleusenstrasse 1, Wilhelmshaven, D-26381, Germany. ¹⁰Department of Integrative Biology, University of Guelph, Guelph, Ontario N1G 2W1, Canada. ¹¹Agricultural and Biosystems Engineering, Iowa State University, Ames, Iowa 50011, USA. ¹²School of Environmental and Forest Sciences, University of Washington, Seattle, Washington 98195, USA. ¹³Department of Biological Sciences, University of Toronto – Scarborough, 1265 Military trail, Toronto, Ontario M1C 1A4, Canada. ¹⁴IFEVA/CONICET – Departamento de Recursos Naturales y Ambiente, Facultad de Agronomía, Universidad de Buenos Aires. Av. San Martín 4453 (C1417DSE) Buenos Aires, Argentina. ¹⁵SYSU-Alberta Joint Lab for Biodiversity Conservation, State Key Laboratory of Biocontrol and School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China. ¹⁶Ecology, Behavior & Evolution Section, University of California, La Jolla, San Diego, California 92093, USA. ¹⁷Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, California 93106-9620 USA. ¹⁸Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado 80309, USA. ¹⁹Department of Entomology, University of Maryland, College Park, Maryland 20742, USA. ²⁰School of Life Sciences, University of KwaZulu-Natal, Pietermaritzburg 3209, South Africa. ²¹School of Biological Sciences, University of Nebraska, Lincoln, Nebraska 68588, USA. ²²Department of Integrative Biology, University of California, Berkeley, California 94720, USA. ²³Department of Plant and Soil Sciences, University of Kentucky, Lexington, Kentucky 40546, USA. ²⁴School of Biological Sciences, Monash University, Victoria 3800, Australia. ²⁵Department of Ecology, Environment and Evolution, La Trobe University, Bundoora, Victoria 3086, Australia. ²⁶CSIRO Land and Water, Private Bag 5, Wembley, Western Australia 6913, Australia. ²⁷Swiss Federal Institute for Forest, Snow and Landscape Research, Community Ecology, Birmensdorf 8903, Switzerland. ²⁸Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK. ²⁹Department of Ecology & Evolutionary Biology, Yale University, 165 Prospect Street, New Haven, Connecticut 06511, USA.

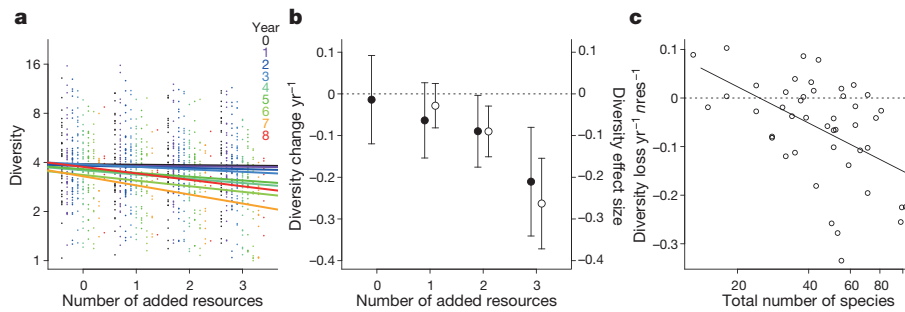


Figure 1 | Biodiversity and number of resources. **a**, Loss of species diversity with greater number of added resources (effective number of equally abundant species: ESN_{PIE}); this effect increased with years of treatment 1–8 (Extended Data Table 1); year 0 shows pre-treatment diversity. Bold lines show overall mean responses of 45 sites; y axis is log-transformed. **b**, Greater number of added resources increased

nutrient addition experiment, replicated across grassland sites on six continents (NutNet; <http://www.nutnet.org>)⁴. We added factorial combinations of phosphorus (P), nitrogen (N), and potassium (K_{+H} ; the K addition treatment included sulfur and a one-time addition of micronutrients; see Methods), with the aim of removing potential limitations from different combinations of the essential nutrient elements that most strongly affect plant growth in natural and managed systems worldwide¹³. Our treatments varied in the number of elemental resources they contained; hereafter, we use the term ‘number of added resources’ (1, 2 or 3) to represent the minimum number of potentially limiting elemental nutrients added (see Methods).

If competition for multiple belowground resources contributes to species coexistence, then diversity should decrease as a function of the number of resources added. Species diversity decreased as more resources were added, and this effect increased with duration of treatment (Fig. 1a and Extended Data Table 1). Greater number of added resources increased the annual rate of diversity loss, even after controlling for differences in experiment duration (Fig. 1b). We found a similar proportional loss of diversity with a greater number of added resources (using the log-ratio effect size of treatment divided by control diversity; Fig. 1b), meaning that in terms of the number of potential species lost, relative diversity losses and annual rate of diversity loss were similar. Sites differed in the size of their species pools, which ranged from 13 to 103 observed species over a three-year period, and we found that the magnitude of diversity loss rate per added resource increased with local species pool size (Fig. 1c).

We found that increasing the number of added resources increased live biomass (Fig. 2a), and decreased the proportion of photosynthetically active radiation (PAR) transmitted through the canopy to the ground surface (Fig. 2b). Further, the amount of litter biomass, which can also contribute to light limitation and diversity loss¹⁴ increased with the number of added resources (Fig. 2c). Importantly, despite the complex causal effects of changes in multiple resources on the

the mean rates of diversity loss per year (filled points; $F_{1,134} = 24.8$, $P < 0.0001$), and the proportional loss of species relative to the controls, shown as the effect size (open points; $F_{1,134} = 46.2$, $P < 0.0001$). **c**, Rate of diversity loss per added resource ($nres$) was associated with greater total site species number (log), $R^2 = 0.25$, $P = 0.0004$, $n = 45$). Error bars show mean \pm 95% confidence intervals.

relationship between diversity and biomass, the number of added resources remained a significant predictor of diversity loss, even after controlling for the potential contributing effects of species pool size, live biomass, total cover (a proxy for total plant abundance), light transmittance, and litter mass (Extended Data Tables 2 and 3). If species coexist though trade-offs in resource-ratio requirements, changes in belowground resource supply could cause changes in competitive dominance and lead to species exclusion², independent of aboveground effects of biomass. In a subset of sites that did not show a biomass response to multiple nutrient addition, we nevertheless observed declines in diversity consistent with this theory (Fig. 3a, b: open points, $n = 11$), similar to sites where biomass production was multiple-resource limited (Fig. 3a, b: filled points, $n = 34$). Overall, 14 sites of 45 sites in this study showed some type of negative biomass response to N, P or K_{+H} addition suggesting the potential for elevated nutrient concentrations supply to cause negative physiological responses in species not adapted to high nutrient concentrations¹⁵ or to large stoichiometric imbalances in resource supply¹⁶.

Diversity loss increased only weakly with biomass increase in plots receiving all three resources, providing some support for indirect effects of biomass as a contributing, but not a sole, mechanism of diversity loss due to fertilisation (Fig. 3c). If species losses were most strongly associated with biomass increases, we would expect the greatest effects on both responses to be associated with the same nutrient addition treatment, but this was true for only 22 of 45 cases (Chi-square, $P < 0.0001$). The loss of diversity was not driven by the addition of any single added resource (for example, N); greatest diversity loss occurred with the addition of a combination of two or more resources in 31 of 45 cases. These findings further highlight that biomass production and diversity can be controlled differently by multiple resources. Overall, these results support our conclusion that resource niche dimensionality can contribute to species diversity independently of indirect effects mediated by biomass production.

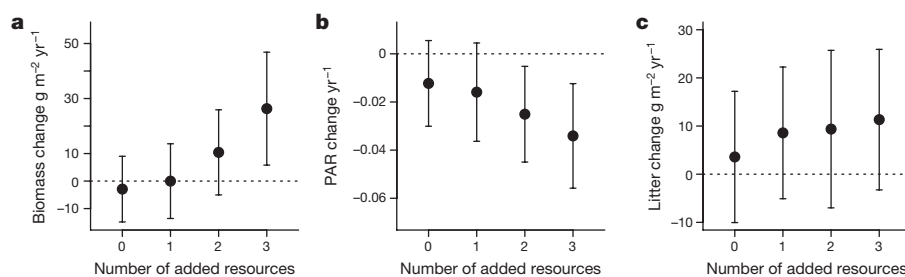


Figure 2 | Biomass and light. **a**, The rate of live biomass change per year increased with an increasing number of added resources ($F_{1,1031} = 55.0$, $P < 0.0001$). **b**, The proportion of photosynthetically active radiation (PAR) reaching the ground surface decreased with a greater number

of added resources, expressed as annual rate of change ($F_{1,782} = 62.4$, $P < 0.0001$). **c**, The mean rate of litter (dead biomass) change per year increased with the number of added resources ($F_{1,783} = 4.37$, $P = 0.037$). Error bars show mean \pm 95% confidence intervals.

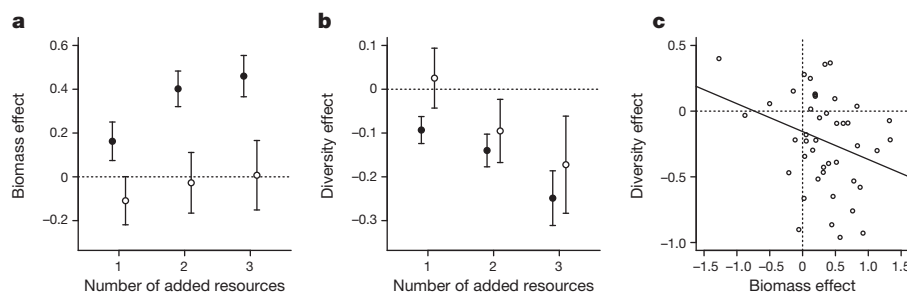


Figure 3 | Multiple resource limitation. **a**, Increased number of added resources resulted in positive and increasing biomass at sites showing multiple resource limitation (filled points); sites not limited by multiple resources tended to show negative biomass responses with resource addition (open points). **b**, Increased number of added resources drove

similar diversity loss at sites where biomass production was limited by multiple resources (filled points) and at sites where it was not (open points). **c**, Negative relationship between the effect of addition of three resources on biomass and diversity (one-tailed test for negative relationship, $R^2 = 0.11$, $P = 0.012$, $n = 45$). Error bars show mean \pm s.e.

For resource dimensionality to contribute to species coexistence, species must trade-off their competitive abilities for different limiting resources, and changes in resource supply ratios should drive species compositional turnover². We found that a greater number of added resources increased the compositional divergence from control plots (Fig. 4a). Plots receiving a single resource treatment (N, P and K_{+M} treatments) diverged as much from each other as they did on average from the control plots (Fig. 4b), consistent with different species trading off competitive abilities for different resources². We found that greater diversity loss was weakly associated with greater community dissimilarity when all three resources were added together (Fig. 4c), suggesting that resource addition caused changes in community composition that were not always associated with diversity loss. Both composition and diversity of communities contribute to ecosystem functioning, and many of the proposed mechanisms of the effect of species diversity on ecosystem function are resource-based¹⁷. Additionally, nutrient enrichment impacts some groups of species more than others (for example, a loss of native species in favour of exotic grasses¹⁸). Because changes in resource supply led to communities of fewer species and of different compositions, we expect changes in resources, acting through diversity loss, to have both direct and indirect effects on ecosystem functions¹⁹.

Although our results are consistent with predictions of the resource niche dimension hypothesis, they are also probably conservative. Our experimental design, a factorial manipulation of three resource treatments, represents a lower-bound estimate of the dimensionality of nutrient resources because our K_{+M} treatment included sulfur and up to 10 other macro- and micro-nutrients, of which more than one may have been limiting¹³. Multiple chemical forms of a limiting nutrient can also contribute to species diversity²⁰, further expanding potential resource dimensionality. Stronger tests of the role of multiple resource competition for structuring species coexistence require

physiological studies quantifying species-specific functional traits and trade-offs²¹, and testing whether species respond to resource treatments similarly in different environments. Deeper mechanistic insight can also be gained by asking how resource-dependent diversity patterns and mechanisms change across scales (for example, from local to regional) in response to global change drivers such as nutrient pollution²². Our results point to, but do not distinguish among, the presumed resource competition mechanisms² that underlie the resource dimension hypothesis.

We found that greater diversity loss was associated with soil P, K, pH and percentage sand, but not with soil N, or with latitude, or mean annual precipitation (Extended Data Table 4), suggesting that variation in soil properties may influence the degree to which communities respond to changes in resource availability²³. We did not test or control for other potential limiting factors such as herbivory or water, which can interact with nutrients in complex ways, and themselves contribute to species coexistence. For example, changes in nutrient availability affect photosynthetic tissue quantity and quality, and may alter the pattern and intensity of herbivory²⁴, and the level of soil water depletion through transpiration losses. Our multi-year experimental results may still under-estimate nutrient effects when considering that global eutrophication represents a chronic and cumulative environmental change over many decades. Estimating effective upper bounds on ecologically relevant resource dimensionality will depend on the degree to which multiple limiting factors covary, how they change in time and space, and how multiple limiting factors interact with each other in promoting coexistence. Global change is driving environmental conditions beyond multiple planetary boundaries²⁵, and changing the limiting factors that structure species diversity²⁶. Understanding the mechanisms that underlie diversity loss caused by multiple global change factors is necessary to develop effective management strategies for restoring and preserving Earth's biodiversity.

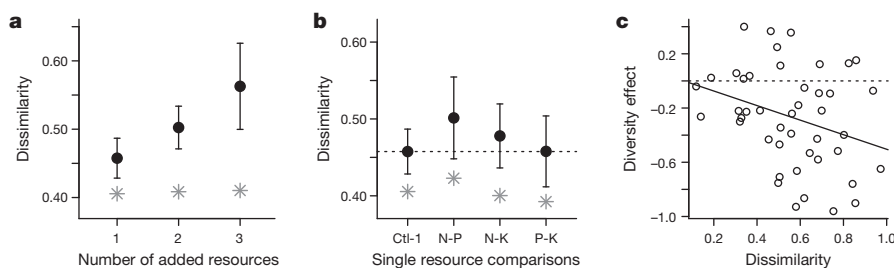


Figure 4 | Community composition. **a**, Community composition diverged from control plots with greater number of added resources (Bray–Curtis dissimilarity index). Resource addition caused greater dissimilarity of community composition relative to mean pre-treatment dissimilarity, indicated by grey stars. **b**, Addition of single nutrient additions of N, P or K_{+M} resulted in communities that diverged as much from each other as

they did on average from the control plots. Pre-treatment values indicated by grey stars. **c**, Negative relationship between the effect of addition of three resources on community dissimilarity relative to controls and diversity (one-tailed test for negative relationship, $R^2 = 0.10$, $P = 0.019$, $n = 45$). Error bars indicate mean \pm 95% confidence intervals.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 February; accepted 21 July 2016.

Published online 24 August 2016.

- Hutchinson, G. E. Concluding remarks. *Quant. Biol.* **22**, 415–427 (1957).
- Tilman, D. *Resource Competition and Community Structure* (Princeton Univ. Press, 1982).
- Hautier, Y., Niklaus, P. A. & Hector, A. Competition for light causes plant biodiversity loss after eutrophication. *Science* **324**, 636–638 (2009).
- Borer, E. T. *et al.* Finding generality in ecology: a model for globally distributed experiments. *Methods Ecol. Evol.* **5**, 65–73 (2014).
- Harpole, W. S. & Tilman, D. Grassland species loss resulting from reduced niche dimension. *Nature* **446**, 791–793 (2007).
- Darwin, C. R. *On the Origin of Species* (John Murray, 1859).
- Interlandi, S. J. & Kilham, S. S. Limiting resources and the regulation of diversity in phytoplankton communities. *Ecology* **82**, 1270–1282 (2001).
- Hutchinson, G. E. The paradox of the plankton. *Am. Nat.* **95**, 137–147 (1961).
- Silvertown, J., Biss, P. M. & Freeland, J. Community genetics: resource addition has opposing effects on genetic and species diversity in a 150-year experiment. *Ecol. Lett.* **12**, 165–170 (2009).
- Ren, Z. *et al.* Effects of resource additions on species richness and ANPP in an alpine meadow community. *J. Plant Ecol.* **3**, 25–31 (2010).
- Harpole, W. S. *et al.* Nutrient co-limitation of primary producer communities. *Ecol. Lett.* **14**, 852–862 (2011).
- Elser, J. J. *et al.* Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. *Ecol. Lett.* **10**, 1135–1142 (2007).
- Fay, P. A. *et al.* Grassland productivity limited by multiple nutrients. *Nature Plants* **1**, 15080 (2015).
- Foster, B. L. & Gross, K. L. Species richness in a successional grassland: effects of nitrogen enrichment and plant litter. *Ecology* **79**, 2593–2602 (1998).
- Chapin, F. S. III The mineral nutrition of wild plants. *Annu. Rev. Ecol. Syst.* **11**, 233–260 (1980).
- Cardinale, B. J., Hillebrand, H., Harpole, W. S., Gross, K. & Ptasnik, R. Separating the influence of resource ‘availability’ from resource ‘imbalance’ on productivity-diversity relationships. *Ecol. Lett.* **12**, 475–487 (2009).
- Tilman, D., Isbell, F. & Cowles, J. M. Biodiversity and ecosystem functioning. *Annu. Rev. Ecol. Syst.* **45**, 471–493 (2014).
- Seabloom, E. W. *et al.* Plant species’ origin predicts dominance and response to nutrient enrichment and herbivores in global grasslands. *Nature Commun.* **6**, 7710 (2015).
- Isbell, F. *et al.* Nutrient enrichment, biodiversity loss, and consequent declines in ecosystem productivity. *Proc. Natl Acad. Sci. USA* **110**, 11911–11916 (2013).
- von Felten, S. *et al.* Belowground nitrogen partitioning in experimental grassland plant communities of varying species richness. *Ecology* **90**, 1389–1399 (2009).
- Litchman, E. & Klausmeier, C. A. Trait-based community ecology of phytoplankton. *Annu. Rev. Ecol. Syst.* **39**, 615–639 (2008).
- Chase, J. M. & Knight, T. M. Scale-dependent effect sizes of ecological drivers on biodiversity: why standardised sampling is not enough. *Ecol. Lett.* **16** (Suppl 1), 17–26 (2013).
- Flores-Moreno, H. *et al.* Climate modifies response of non-native and native species richness to nutrient enrichment. *Phil. Trans. R. Soc. B* **371**, 20150273 (2016).
- Borer, E. T. *et al.* Herbivores and nutrients control grassland plant diversity via light limitation. *Nature* **508**, 517–520 (2014).
- Steffen, W. *et al.* Planetary boundaries: guiding human development on a changing planet. *Science* **347**, 1259855 (2015).
- Tilman, D. & Lehman, C. Human-caused environmental change: impacts on plant diversity and evolution. *Proc. Natl Acad. Sci. USA* **98**, 5433–5440 (2001).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the Minnesota Supercomputer Institute for hosting project data, the University of Minnesota Institute on the Environment for hosting Nutrient Network meetings, and each site investigator for funding their site-level operations. Network coordination and data management were supported by funds from the National Science Foundation Research Coordination Network (NSF-DEB-1042132) to E.T.B. and E.W.S. from the Long Term Ecological Research program (NSF-DEB-1234162) to the Cedar Creek LTER, and from the Institute on the Environment (DG-0001-13). Konza NutNet site was funded by the Konza Prairie LTER; the Saline Experimental Range NutNet site was funded by a Yale Institute for Biospheric Studies Pilot Grant. Nitrogen fertilizer was donated to the Nutrient Network by Crop Production Services, Loveland, Colorado. We thank N. Gotelli for discussions.

Author Contributions W.S.H. analysed the data and wrote the paper with contributions and input from all authors. L.L.S., E.M.L. and J.F. contributed to data analysis. W.S.H., E.W.S. and E.T.B. developed and framed the research questions. W.S.H., E.W.S., E.T.B. and E.M.L. are Nutrient Network coordinators. All authors collected data used in this analysis. Author contribution matrix provided as Supplementary Table 2.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to W.S.H. (stan.harpole@idiv.de).

Reviewer Information *Nature* thanks J. Levine, B. Schmid and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The investigators were not blinded to allocation during experiments and outcome assessment.

Experimental design. The Nutrient Network (NutNet) is a collaborative, distributed experimental network⁴. Sites are located across herbaceous terrestrial systems on six continents. Vegetation types represented include grasslands, savannas and meadows and occur across a wide range of climate and environmental factors (Supplementary Table 1). At the 45 sites (on five continents) with appropriate experimental data for these analyses, one year of pre-treatment (year 0) data were collected followed by at least 3 years and up to 8 years of treatment data. Individual site experiments share identical design and sampling protocols, with minor site-specific differences in terms of replication and treatment duration (Supplementary Table 1). We applied factorial combinations of nitrogen (N), phosphorus (P), and potassium plus micronutrients, designated here as the $K_{+μ}$ treatment, giving eight treatment combinations including the control with no added resources. N was applied annually at $10\text{ g N m}^{-2}\text{ yr}^{-1}$ as time-release urea. Ammonium nitrate was used in 2007 at some sites before switching to urea due to restricted availability of ammonium nitrate; we found no differences in the short-term effects of alternative N sources in a separate experiment at four sites¹⁸. P was applied at $10\text{ g P m}^{-2}\text{ yr}^{-1}$ as triple-super phosphate, which also included Ca at $8.1\text{ g Ca m}^{-2}\text{ yr}^{-1}$. The $K_{+μ}$ treatment added a mix of K and S ($10\text{ g K m}^{-2}\text{ yr}^{-1}$ and $3.9\text{ g S m}^{-2}\text{ yr}^{-1}$ as potassium sulphate) and micronutrients ($100\text{ g m}^{-2}\text{ yr}^{-1}$ of a mixture composed of 6% Ca, 3% Mg, 12% S, 0.1% B, 1% Cu, 17% Fe, 2.5% Mn, 0.05% Mo, and 1% Zn). Micronutrients were only applied during the first treatment year to minimise potential for toxic metal accumulation. Plots were $5\text{ m} \times 5\text{ m}$ and randomized within 1 to 6 blocks (Supplementary Table 1), with all eight treatment combinations occurring once per block. Sampling occurred at approximately peak biomass times for each site.

Response measurements. Biodiversity estimates are scale-dependent²⁷, and increased resource availability can alter diversity-scaling relationships by changing the size of species pools and thus introduce differences in the coverage of sampling between treatments, due to larger and fewer individuals per area sampled²⁸, and contribute to the loss of rarer species. We calculated species diversity as the effective species number, which estimates the probability of interspecific-encounter if all species are equally abundant (ESN_{PIE}). ESN_{PIE} has been shown to be less sensitive to scaling issues than other metrics²², and is representative of the maximum slope of the species-area accumulation function. We used ESN_{PIE} because NutNet sites vary in their species pools and therefore their species accumulation curves will differ, which creates a challenge to compare species diversity when sampled at a fixed area²². ESN_{PIE} has been shown to be relatively insensitive to such sampling area issues because it essentially measures the maximum change in species number as a function of sampling area (that is, the slope at the x intercept of the species accumulation curve). Because the resource dimension hypothesis and underlying resource ratio theory assume that species trade-off for different limiting factors, predictions for diversity change describe changes in competitive dominance; ESN_{PIE} captures these predicted changes in dominance better than simple measurements of local species extinction (that is, richness loss). We used the aggregate number of species observed at a site as an estimate of the asymptote of the species accumulation function, and of the regional species pool. We also used simply the number of species (that is, richness) and found similar results as those using ESN_{PIE} (Extended Data Table 1).

We measured species diversity annually by estimating the % cover of each plant species within a $1\text{ m} \times 1\text{ m}$ fixed location in each plot; the total cover typically summed to greater than 100% due to multiple canopy layers. We quantified species diversity as the probability of interspecific encounter (PIE), or effective species number (ESN_{PIE}), assuming species relative abundances are equal:

$$ESN_{PIE} = \frac{1}{\sum_i^s p_i^2} \quad (1)$$

where p_i is the proportion of species i in a community of size s ; ESN_{PIE} is derived from the inverse of Simpson's diversity index²².

We measured aboveground live biomass by clipping two $1\text{ m} \times 10\text{ cm}$ strips of vegetation in each plot, sorting the sampled tissue to live (current year's production) and dead (previous years' production) fractions, drying at 60°C for 48 h and weighing. At most sites, photosynthetically active radiation PAR was measured above the plant canopy and at the ground surface and the proportion of transmitted light calculated.

We categorised plant communities at sites as multiple-resource limited if biomass responded positively to fertilisation with combinations of different nutrients. Specifically, we designated sites as 'multiple-resource limited' if biomass increased with the independent addition of different resources or if biomass responded synergistically to two or more added resources (that is, the response to one nutrient was dependent on the level of another and their combined effect was super-additive)¹¹. Sites that showed no response or negative biomass response or responded positively to only one resource we categorised as not multiple-resource limited. Thirty-four of the 45 sites showed increased biomass in response to multiple added resources; eight did not respond positively to resource addition, and three responded positively to a single resource (that is, single resource limited¹¹).

Statistical analysis. All analyses used R version 3.2.2. We used linear mixed-effects models (R package lme) to test the interaction of number of added resources and the number of treatment years, on diversity (ESN_{PIE}) and richness. Site and block were modelled as nested random effects. We included in the model an autocorrelation structure, a first-order autoregressive model (AR(1)), where observations are expected to be correlated from one year to the next, and found a substantial improvement in model fit when we compared this model to a model with no autocorrelation structure (lower AIC = $\Delta 608$ and likelihood ratio tests, $L.Ratio = 610$, $P < 0.0001$)²⁹. Treatment effects increased in magnitude with time (significantly negative interaction between number of added resources and year; Supplementary Table 2). To allow standardized comparison of sites that differed in the year they were established and in duration of nutrient addition, we used two approaches to quantify the changes in species diversity. First, we calculated the annual rate of change of our response variables to standardise site responses. Second, for analyses that required an effect size, calculated as the log ratio of the treatment response divided by the control, we used the most recent year of treatment data, which ranged from 3 to 8 years of annual nutrient application duration (Supplementary Table 1). Log ratio effect size estimates would not have been possible using the rate of change estimates, which can take zero or negative values. Log ratio effect sizes tend to be normally distributed, centre zero effects (control levels) at zero log ratios, and scale responses to make proportional effects directly comparable between sites³⁰.

We used linear mixed-effects models (R package lme) to test the effects of number of treatment years, site richness, log live biomass, log dead biomass, PAR, total species cover, and the number of added resources on diversity (ESN_{PIE}), with plot nested in block nested in site as random effects. Models using dead biomass and PAR used the subset of 32 sites for which we had data for these variables. We calculated mean values at each site for the annual rate of diversity loss and diversity effect size, and tested for linear relationships between these variables and the number of added resources using regression with site as a block term. We used step-wise linear regression and AIC criteria to test for relationships of loss of diversity (from addition of three resources) with latitude, longitude, and environmental covariates of mean annual precipitation, and soil N, P, K, pH, percentage clay, and percentage sand. Plant community composition changes were quantified using Bray-Curtis multivariate distances (R package vegan).

27. Crawley, M. J. & Harral, J. E. Scale dependence in plant biodiversity. *Science* **291**, 864–868 (2001).
28. Oksanen, J. Is the humped relationship between species richness and biomass an artefact due to plot size? *J. Ecol.* **84**, 293–295 (1996).
29. Pinheiro, J. & Bates, D. *Mixed-effects models in S and S-PLUS* (Springer, 2006).
30. Hedges, L. V., Gurevitch, J. & Curtis, P. S. The meta-analysis of response ratios in experimental ecology. *Ecology* **80**, 1150–1156 (1999).

Extended Data Table 1 | The effects of nutrient addition on diversity loss and richness loss increase with time

ESN_{pie}	Num. DF	Den. DF	F	P
intercept	1	6555	370.3	<0.0001
year	1	6555	39.3	<0.0001
nres	1	1049	32.2	<0.0001
year x nres	1	6555	26.1	<0.0001

Richness	Num. DF	value	SE	P
intercept	1	6555	146.4	<0.0001
year	1	6555	209.1	<0.0001
res	1	1049	91.8	<0.0001
year x nres	1	6555	33.5	<0.0001

Linear mixed-effects model of the effects of number of treatment years (ARIMA type-1 autocorrelation) and the number of added resources on diversity (log ESN_{pie}) and richness, with plot nested in block, nested in year, nested in site, as random effects, using all 45 sites. There was a significant, negative interaction between the number of added resources (*nres*) and year of treatment (*year*).

Extended Data Table 2 | The number of added resources predicts diversity loss after controlling for other variables

	Num. DF	Den. DF	F	P
intercept	1	1029	329.9	<0.0001
years of treatment	1	42	12.1	0.0012
site richness	1	42	21.5	<0.0001
log live biomass	1	1029	17.4	<0.0001
total cover	1	1029	4.8	0.029
number of added resources	1	1029	35.4	<0.0001

Linear mixed-effects model of the effects of number of treatment years, site richness, log live biomass, total species cover, and the number of added resources on diversity (ESN_{pie}), with plot nested in block nested in site as random effects, using all 45 sites and data from the maximum treatment year for each site. Δ AIC between model with number of added resources and model without was 33, log-likelihood ratio 35.0, $P < 0.0001$.

Extended Data Table 3 | The number of added resources is an important predictor even after controlling for other variables, for sites that had light and litter data

	Num. DF	Den. DF	F	P
intercept	1	643	285.3	<0.0001
years of treatment	1	29	14.1	0.0008
site richness	1	29	25.7	<0.0001
log live biomass	1	643	7.9	0.0052
total cover	1	643	4.5	0.034
log dead biomass	1	643	0.34	0.56
PAR	1	643	18.2	<0.0001
number of added resources	1	643	15.6	0.0001

Linear mixed-effects model of the effects of number of treatment years, site richness, log live biomass, log dead biomass, PAR, total species cover, and the number of added resources on diversity (ESN_{plu}), with plot nested in block nested in site as random effects, using data from the maximum treatment year for each site, and the subset of 32 sites for which there was dead biomass and PAR data. Δ AIC between model with number of added resources and model without was 15, log-likelihood ratio 15.6, $P < 0.0001$.

Extended Data Table 4 | Diversity loss due to addition of nutrients associated with soil properties

	DF	SS	F	P
soil P	1	0.16	1.72	0.20
soil K	1	0.018	0.20	0.66
pH	1	0.46	5.03	0.034
% sand	1	0.73	8.05	0.0089
residuals	25	2.28		

Stepwise multiple regression (backward with AIC criteria for model comparisons) retained soil P, K, pH, and percentage sand as predictors of diversity loss from the addition of three resources, for the 30 sites with soil analysis data (excluding one site for extreme value of P). The variables latitude, longitude, mean annual precipitation, and soil percentage N were not retained. Overall model is significant ($r^2 = 0.375$, $F_{4,25} = 3.75$, $P = 0.016$).

Serotonin engages an anxiety and fear-promoting circuit in the extended amygdala

Catherine A. Marcinkiewicz^{1*}, Christopher M. Mazzone^{1,2*}, Giuseppe D'Agostino³, Lindsay R. Halladay⁴, J. Andrew Hardaway¹, Jeffrey F. DiBerto¹, Montserrat Navarro⁵, Nathan Burnham⁵, Claudia Cristiano³, Cayce E. Dorrier¹, Gregory J. Tipton¹, Charu Ramakrishnan⁶, Tamas Kozicz^{7,8}, Karl Deisseroth⁶, Todd E. Thiele^{1,5}, Zoe A. McElligott^{1,9}, Andrew Holmes⁴, Lora K. Heisler³ & Thomas L. Kash^{1,2,5,10}

Serotonin (also known as 5-hydroxytryptamine (5-HT)) is a neurotransmitter that has an essential role in the regulation of emotion. However, the precise circuits have not yet been defined through which aversive states are orchestrated by 5-HT. Here we show that 5-HT from the dorsal raphe nucleus (5-HT^{DRN}) enhances fear and anxiety and activates a subpopulation of corticotropin-releasing factor (CRF) neurons in the bed nucleus of the stria terminalis (CRF^{BNST}) in mice. Specifically, 5-HT^{DRN} projections to the BNST, via actions at 5-HT_{2C} receptors (5-HT_{2C}Rs), engage a CRF^{BNST} inhibitory microcircuit that silences anxiolytic BNST outputs to the ventral tegmental area and lateral hypothalamus. Furthermore, we demonstrate that this CRF^{BNST} inhibitory circuit underlies aversive behaviour following acute exposure to selective serotonin reuptake inhibitors (SSRIs). This early aversive effect is mediated via the corticotrophin-releasing factor type 1 receptor (CRF₁R, also known as CRHR1), given that CRF₁R antagonism is sufficient to prevent acute SSRI-induced enhancements in aversive learning. These results reveal an essential 5-HT^{DRN} → CRF^{BNST} circuit governing fear and anxiety, and provide a potential mechanistic explanation for the clinical observation of early adverse events to SSRI treatment in some patients with anxiety disorders^{1,2}.

Give the multiple converging lines of evidence pinpointing 5-HT as a critical neuromodulator of pathological fear learning^{3,4}, we first interrogated the endogenous recruitment of the 5-HT^{DRN→BNST} circuit by an aversive footshock stimulus in mice. Using Fluoro-Gold to retrogradely label BNST-projecting 5-HT neurons in the dorsal raphe nucleus (DRN), we found that *c-fos*, an immediate-early gene indicative of *in vivo* neuronal activation, was significantly elevated in 5-HT^{DRN→BNST} neurons after footshock (Fig. 1a–f). Using *in vivo* electrophysiology, we then probed the neuronal dynamics of the BNST during fear conditioning and recall, and found evidence for engagement during both conditioning and recall (Extended Data Fig. 1).

To decipher the role of this 5-HT^{DRN→BNST} circuit in aversive behaviour, Channelrhodopsin2 (ChR2)–eYFP was selectively expressed in 5-HT^{DRN} neurons through the delivery of a Cre-inducible viral vector in mice expressing Cre recombinase under the control of a serotonin transporter promoter (*Sert*^{Cre}) (*Sert* is also known as *Slc6a4*) for both *in vivo* and *ex vivo* analysis. We observed eYFP⁺ (5-HT) cell bodies in the DRN and eYFP⁺ fibres in both the dorsal and ventral aspects of the BNST (*Sert*^{Cre}::ChR2^{DRN→BNST}), confirming a direct projection of 5-HT neurons originating in the DRN to the BNST (Fig. 1g, h)⁵. Optical stimulation of these fibres in BNST slices evoked 5-HT release, as measured by fast-scan cyclic voltammetry (FSCV) (Fig. 1i, j). Furthermore, bath

application of the SSRI fluoxetine reliably decreased the rate of 5-HT reuptake, confirming that photostimulation of SERT⁺ terminals in the BNST originating from the DRN induces 5-HT release (Fig. 1k, l).

We examined whether this 5-HT^{DRN→BNST} circuit is functionally relevant for fear and anxiety-like behaviour. To investigate this, *Sert*^{Cre}::ChR2^{DRN→BNST} mice were implanted with bilateral optical fibres and photostimulated in the BNST (473 nm, 20 Hz) using a standard tone-shock fear conditioning paradigm. Optogenetic stimulation of this pathway was paired with a tone that co-terminated with a scrambled footshock. Cued fear was assessed 24 h after, and contextual fear 48 h after, the initial fear acquisition session (Fig. 1m, n). Although no changes were observed during fear acquisition, both cued and contextual fear recall were significantly heightened in photostimulated *Sert*^{Cre}::ChR2^{DRN→BNST} mice (Fig. 1o–q). We next assessed anxiety-like behaviour using well-characterized assays: the elevated plus maze (EPM) and novelty-suppressed feeding (NSF) tests. Upon stimulation with light, *Sert*^{Cre}::ChR2^{DRN→BNST} mice exhibited enhanced anxiety-like behaviour in both the EPM and NSF tests (Fig. 1r, s and Extended Data Fig. 2a, b). Importantly, photostimulation did not induce hypolocomotion in the EPM or open field tests, nor did it alter home-cage feeding, thus confirming that hypophagia in the NSF assay was due to anxiety and not a reduction in appetitive drive (Extended Data Fig. 2c–e). One potential explanation of these results is that terminal stimulation in the BNST produces antidromic spikes in DRN cell bodies that release 5-HT in other brain regions, which could be also be driving these behaviours. Therefore we probed the mechanism more deeply using converging approaches.

To determine a receptor target through which 5-HT is signalling in the BNST, we then examined the impact of optogenetically evoked 5-HT^{DRN} release on postsynaptic neuronal excitability and found a 3.05 ± 0.59 mV depolarization that was blocked by a 5-HT_{2C}R antagonist (Fig. 1t, u). In contrast to previous reports demonstrating co-release of 5-HT and glutamate from DRN projections to the nucleus accumbens⁶, we did not observe any time-locked light-evoked EPSCs in the BNST (data not shown). These results indicate that 5-HT^{DRN→BNST} projections have a predominantly excitatory effect that is dependent on 5-HT_{2C}R signalling. To examine the role of 5-HT_{2C}R-containing neurons in anxiety-like behaviour, we took advantage of a *Htr2c*^{Cre} mouse line (Extended Data Fig. 3a, b)⁷. Using 'designer receptors exclusively activated by designer drugs' (DREADDs) that are coupled to the G_{oq} signalling pathway (hM3Dq-DREADD)⁸, we found that activation of G_{oq} signalling in 5-HT_{2C}R-expressing neurons in the BNST significantly delayed the onset of feeding in the NSF assay without affecting home cage feeding behaviour (Extended Data Fig. 3c–g), thus phenocopying the effect observed

¹Bowles Center for Alcohol Studies, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. ²Curriculum in Neurobiology, School of Medicine, University of North Carolina at Chapel Hill School of Medicine, Chapel Hill, North Carolina 27599, USA. ³Rowett Institute of Nutrition and Health, University of Aberdeen, Aberdeen AB25 2ZD, UK. ⁴National Institute on Alcohol Abuse and Alcoholism, National Institutes of Health, Rockville, Maryland 20852-9411, USA. ⁵Department of Psychology & Neuroscience, College of Arts and Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. ⁶Department of Bioengineering, Stanford University, Stanford, California 94305, USA. ⁷Hayward Genetics Center, Tulane University, New Orleans, Louisiana 70112, USA. ⁸Department of Anatomy, Radboud University Nijmegen Medical Center, 6500HB Nijmegen, The Netherlands. ⁹Department of Psychiatry, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. ¹⁰Department of Pharmacology, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA.

*These authors contributed equally to this work.

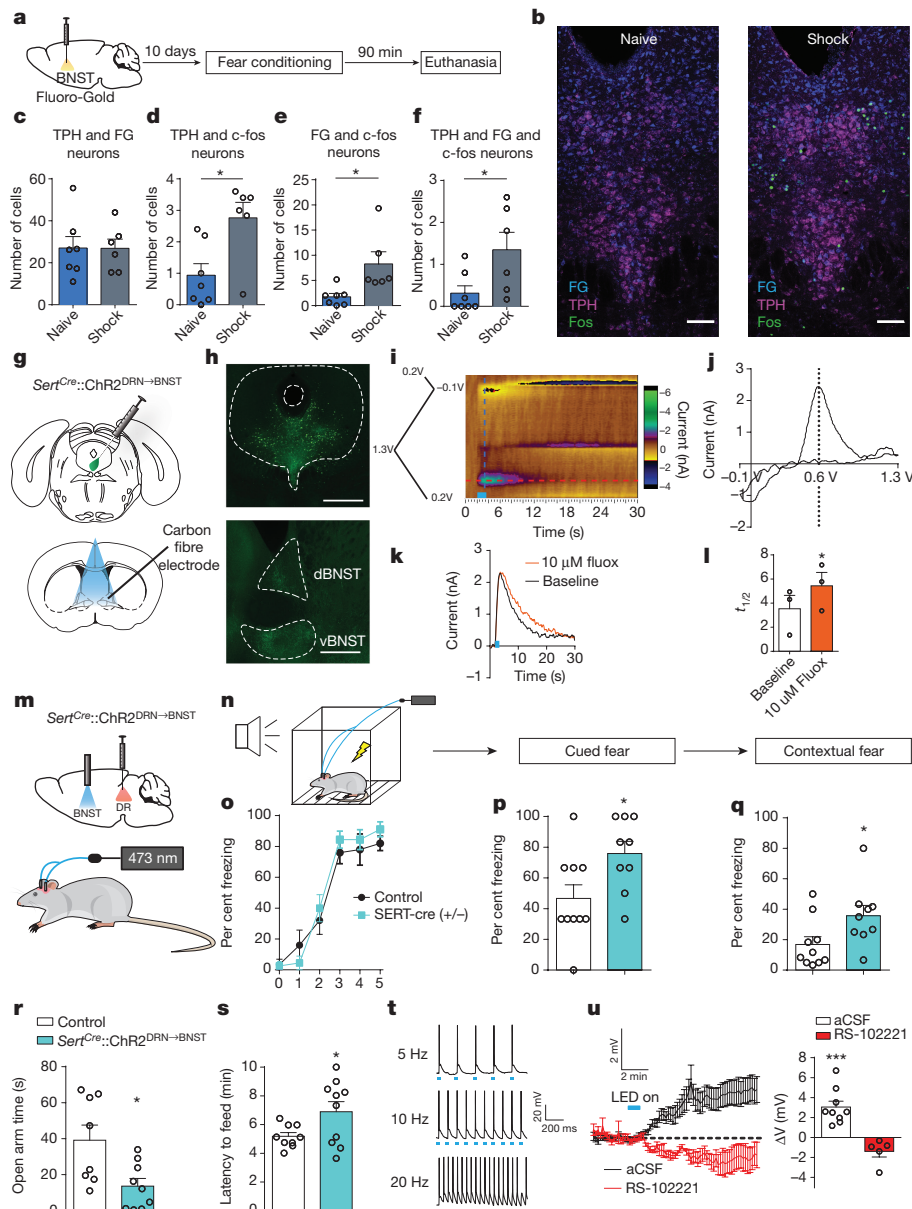


Figure 1 | Optogenetic identification of a 5-HT^{DRN→BNST} projection that elicits anxiety and fear-related behaviour. **a**, Experimental timeline for c-fos labelling of 5-HT^{DRN→BNST} neurons following an aversive footshock stimulus. **b**, Representative images of Fluoro-Gold (FG, blue), tryptophan hydroxylase (TPH, violet), and c-fos (green) staining in the DRN for 13 mice. Scale bars, 100 μ m. **c–f**, Histograms depicting the number of double- and triple-labelled neurons in the DRN of naive and shocked mice. **c**, There were no significant differences in the number of BNST-projecting 5-HT^{DRN} neurons between groups. **d–f**, Footshock lead to significant elevations in the number of c-fos⁺ 'activated' 5-HT neurons ($t_{11} = 2.975$, $P < 0.05$, Student's unpaired two-tailed t -test, $n = 7$ naive and $n = 6$ shocked mice), c-fos⁺, Fluoro-Gold-labelled neurons ($t_{11} = 2.836$, $P < 0.05$, Student's unpaired two-tailed t -test, $n = 7$ naive and $n = 6$ shocked mice), and triple-labelled neurons ($t_{11} = 2.374$, $P < 0.05$, Student's unpaired two-tailed t -test, $n = 7$ naive and $n = 6$ shocked mice). **g**, Experimental configuration for light-evoked FSCV experiments in *Sert*^{Cre}:ChR2^{DRN→BNST} mice. **h**, Coronal images showing ChR2-YFP expression in the soma of the DRN and axons of the BNST. Scale bars, 500 μ m. **i**, Representative colour plot of 5-HT release to optical stimulation (blue bar, 20 Hz, 20 pulses) for 3 mice. **j**, Representative cyclic voltammogram at peak 5-HT (black dashed line) for 3 mice. **k**, Representative current versus time

trace at baseline (black) and following 10 μ M fluoxetine (red) for 3 mice. **l**, Clearance half-life of 5-HT at baseline (white) and following 10 μ M fluoxetine (red). $t_2 = 8.43$, $P < 0.05$, Student's paired two-tailed t -test, $n = 3$ slices from 3 mice. **m**, *Sert*^{Cre} mice were transduced in the DRN and implanted with bilateral optical fibres in the BNST. **n**, Schematic of fear conditioning procedures in *Sert*^{Cre}:ChR2^{DRN→BNST} mice. **o–q**, Photostimulation during fear acquisition had no effect on freezing behaviour during fear learning but increased freezing during cued ($t_{17} = 2.436$, $P < 0.05$, Student's unpaired two-tailed t -test, $n = 10$ control, $n = 9$ ChR2) and contextual fear recall ($t_{17} = 2.271$, $P < 0.05$, Student's unpaired two-tailed t -test, $n = 10$ control, $n = 9$ ChR2). **r**, Light delivery to the BNST reduced open arm time in the EPM ($t_{15} = 2.79$, $P < 0.05$, Student's unpaired two-tailed t -test, $n = 8$ control, $n = 9$ ChR2). **s**, Increased latency to feed in the NSF ($t_{17} = 2.19$, $P < 0.05$, Student's unpaired two-tailed t -test, $n = 9$ control, $n = 10$ ChR2). **t**, Action potentials generated by photostimulation in the DRN (5 Hz (top), 10 Hz (middle), 20 Hz (bottom), 473 nm). **u**, Depolarization in cells ($t_8 = 5.20$, $P < 0.01$, one-sample t -test, $n = 9$ cells from 4 mice) after photostimulation in the BNST (5 Hz, 10 s, 473 nm) and blockade of this response by 5 μ M RS-102221 (5-HT_{2C}R antagonist) ($t_4 = 2.5$, $P > 0.05$, one-sample t -test, $n = 5$ cells from 2 mice). Data are mean \pm s.e.m. * $P < 0.05$; *** $P < 0.001$.

with 5-HT^{DRN→BNST} fibre stimulation during NSF. Together, these results provide converging evidence that activation of 5-HT^{DRN→BNST} inputs elicits anxiety-like behaviour via 5-HT_{2C}R signalling.

We considered the neurochemical phenotype of these target 5-HT^{DRN→5-HT_{2C}R^{BNST}} neurons and hypothesized that 5-HT via 5-HT_{2C}R modulates the activity of neurons expressing the neuropeptide

CRF. This hypothesis was based on a previous analysis of 5-HT_{2C}R knockout mice, which exhibit an anxiolytic phenotype associated with a reduction of c-fos in CRF^{BNST} neurons⁹. Initially, using CRF reporter mice to *a priori* select CRF neurons for recordings, we found a heterogeneous 5-HT-induced response in CRF^{BNST} neurons (Extended Data Fig. 4a), with only a subset demonstrating a depolarization. Consistent with this, double fluorescence *in situ* hybridization revealed that only a subset of CRF neurons within the dorsal BNST (~70%) and ventral BNST (~43%) express 5-HT_{2C}R (Extended Data Fig. 4b–d).

Although CRF signalling within the BNST is associated with anxiety-like behaviour^{10,11}, more recent studies using circuit-based tools have found that optogenetic stimulation of GABAergic projections (which include CRF^{BNST} neurons) to the ventral tegmental area (VTA) are anxiolytic¹². This led us to hypothesize the existence of functionally distinct subsets of CRF^{BNST} neurons that gate different behaviours and are differentially sensitive to 5-HT. We used fluorescent retrograde tracer beads to label CRF^{BNST} neurons as VTA-projecting or non-VTA-projecting (Fig. 2a), and found that VTA-projecting CRF neurons (CRF^{BNST→VTA} neurons) were hyperpolarized by an average of 5.73 ± 1.24 mV and non-VTA-projecting CRF neurons were depolarized by an average of 2.74 ± 0.39 mV during 5-HT bath application. Moreover, the excitatory response to 5-HT in non-VTA-projecting CRF neurons was reversed in the presence of a 5-HT_{2C}R antagonist (Fig. 2b). Furthermore, all CRF^{BNST→VTA} neurons were non-responsive to the 5-HT_{2C}R agonist meta-chlorophenylpiperazine (mCPP), whereas all non-VTA projecting CRF neurons were depolarized by mCPP by an average of 3.26 ± 0.74 mV (Extended Data Fig. 4e–h). These findings suggest an anatomically distinct response to 5-HT by different subsets of CRF^{BNST} neurons. The subset of CRF^{BNST} neurons expressing

5-HT_{2C}R do not project to the VTA and are depolarized by 5-HT, whereas the CRF^{BNST→VTA} neurons are hyperpolarized by 5-HT, via actions at another 5-HT receptor.

To determine if this 5-HT-dependent mechanism extended to other anxiolytic efferents, we injected retrograde tracer beads into the lateral hypothalamus (LH) of CRF reporter mice and found 5-HT had similar bidirectional effects on non-LH-projecting and LH-projecting CRF^{BNST} neurons (Extended Data Fig. 5a–c). Noting the functional similarities between these two populations, we used retrograde tracing to determine that roughly ~58% of CRF^{BNST} neurons have projections to the LH or VTA (Extended Data Fig. 5d–f). Notably, ~20–31% of these CRF^{BNST} output neurons form parallel projections to these structures.

In light of recent reports that CRF^{BNST} neurons are exclusively GABAergic¹³, we hypothesized that non-VTA-projecting CRF^{BNST} neurons may locally inhibit BNST→VTA neurons to promote fear and anxiety. To test this hypothesis, we injected *Crf*^{Cre} mice with a Cre-inducible ChR2 into the BNST and retrograde tracer beads into the VTA. We then recorded light-evoked inhibitory postsynaptic potentials (IPSCs) from non-ChR2 (ChR2-negative, retrograde tracer-positive) VTA-projecting BNST neurons (Fig. 2c). Photostimulation produced action potentials in CRF^{BNST} neurons and light-evoked IPSCs in non-ChR2 VTA-projecting neurons, indicating that CRF^{BNST} neurons form local GABAergic synapses with BNST neurons that project to the VTA. Repeating these same experiments in *Crf*^{Cre::ChR2} mice with retrograde tracer beads in the LH, we found that we could evoke GABA currents using photostimulation in LH-projecting neurons as well (Extended Data Fig. 5g–i). Moreover, we observed that 5-HT increased GABAergic transmission on to BNST→VTA projecting neurons in a tetrodotoxin and 5-HT_{2C}R antagonist dependent manner (Fig. 2d–f and Extended

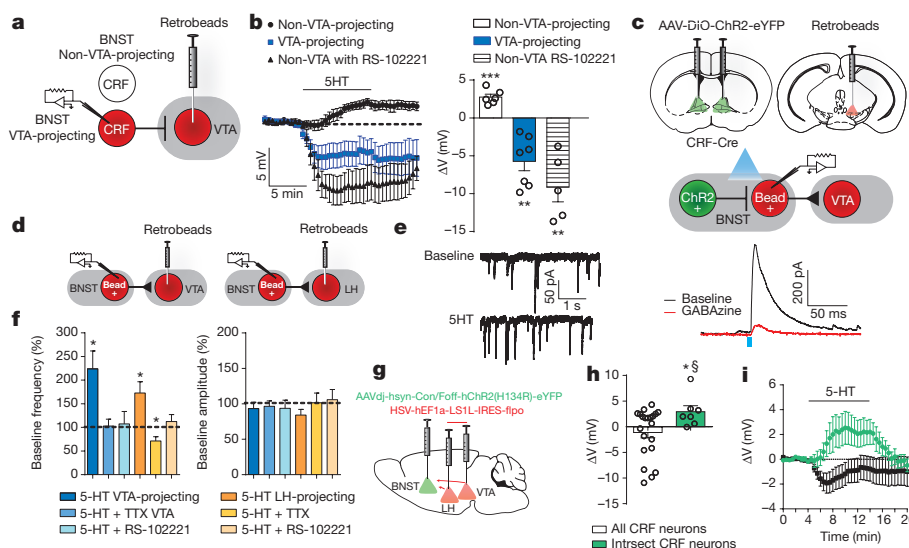


Figure 2 | Serotonin activates a local population of CRF^{BNST} neurons that inhibits outputs to the midbrain. **a**, Recording scheme for CRF reporter mice injected with retrograde tracer beads in the VTA. **b**, 5-HT depolarizes local CRF neurons ($t_5 = 7.06$, $P < 0.001$, one-sample *t*-test, $n = 6$ cells from 4 mice) in the BNST while hyperpolarizing CRF^{BNST→VTA} neurons ($t_6 = 4.64$, $P < 0.01$, one-sample *t*-test, $n = 7$ cells from 6 mice). Non-VTA-projecting CRF neurons are hyperpolarized by 5-HT in the presence of the 5-HT_{2C}R antagonist RS-102221 ($t_4 = 4.74$, $P < 0.01$, one-sample *t*-test, $n = 5$ cells from 3 mice). **c**, Top and middle, schematic depicting infusions and recording configuration for *Crf*^{Cre}:ChR2^{BNST} mice injected with retrograde tracer beads in the VTA. Bottom, representative trace of light-evoked IPSC in beaded (that is, VTA projecting), non-ChR2 expressing neurons in the BNST of *Crf*^{Cre}:ChR2 mice with retrograde tracer beads in the VTA ($n = 8$ cells from 3 mice) and blockade of this response by GABAzine ($F_{1,33} = 53.16$, $P < 0.001$, repeated measures one-way ANOVA, $n = 4$ cells from 3 mice). **d**, Recording scheme for C57BL/6 mice with retrograde tracer beads in the VTA or LH. **e**, Representative

traces of sIPSCs in BNST neurons that project to the VTA before and after 5-HT application for 5 cells from 4 mice. **f**, Bar graphs showing magnitude of 5-HT effect on average sIPSC frequency in BNST neurons that project to the VTA ($t_4 = 3.257$, $P < 0.05$, one-sample *t*-test, $n = 5$ cells from 4 mice) and in BNST neurons that project to the LH ($t_5 = 3.027$, $P < 0.05$, one-sample *t*-test, $n = 6$ cells from 3 mice) and blockade of these responses by tetrodotoxin (TTX) and RS-102221. Effects on amplitude were non-significant. **g**, Experimental scheme for experiments with *Crf*^{Cre::Intersc}:ChR2^{BNST} mice. **h**, 5-HT significantly depolarizes non-projecting CRF (Intersc) neurons in the BNST ($t_6 = 2.501$, $P < 0.05$, one-sample *t*-test, $n = 7$ cells from 5 mice) and produces a significant change in membrane potential in CRF Intersc neurons compared to all CRF neurons ($t_{26} = 2.08$, $P < 0.05$, Student's unpaired two-tailed *t*-test, $n = 21$ cells from 14 mice for experiments in all CRF neurons and $n = 7$ cells from 5 mice for *Crf*^{Cre::Intersc}:ChR2^{BNST} experiments). Data are mean \pm s.e.m. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$. § denotes $P < 0.05$ for the Student's unpaired two-tailed *t*-test between all CRF neurons and CRF Intersc neurons in **h**.

Data Fig. 5j–n). Similar effects of 5-HT on GABAergic transmission were found in BNST→LH projecting neurons (Extended Data Fig. 5o–v). Furthermore, slice recordings in a CRF reporter line indicates that 5-HT does not increase GABAergic transmission on to the general population of CRF^{BNST} neurons nor does it directly excite non-CRF VTA projecting neurons (Extended Data Fig. 6). The 5-HT₂R agonist mCPP also increased GABAergic but not glutamatergic transmission in the BNST (Extended Data Fig. 7). Finally, to test if optically evoked 5-HT can inhibit BNST outputs to the VTA, we performed slice recordings in the BNST of *Sert^{Cre}::Chr2^{DRN→BNST}* mice and found that brief photostimulation of 5-HT terminals in the BNST increased spontaneous IPSCs (sIPSCs) on to VTA projecting BNST neurons in a manner similar to bath-applied 5-HT (Extended Data Fig. 8a–c). Together, these experiments indicate that CRF^{BNST} neurons inhibit at least two major BNST outputs to the VTA and LH that are reported to be anxiolytic^{12,14}, providing mechanistic insight into the aversive actions of 5-HT signalling in the BNST.

We took advantage of a new combinatorial strategy called INTronic Recombinase Sites Enabling Combinatorial Targeting or INTRASECT¹⁵ that allows for direct visualization of these non-projecting, putatively local CRF^{BNST} neurons in the BNST. By coupling retrograde Cre-dependent flippases (HSV-LSL1-mCherry-IRES-flpo) in the VTA and LH with a (*Cre_{on}/flp_{off}*)-Chr2-eYFP viral construct in the BNST of *Cr^fCre* mice (*Cr^fCre::Intrsect-Chr2^{BNST}* mice), we were able to genetically isolate

non-VTA/LH-projecting CRF neurons in the BNST. We also infused a Cre-dependent HSV-mCherry vector in a subset of *Cr^fCre::Intrsect-Chr2^{BNST}* mice as a control. In HSV-flpo infused *Cr^fCre::Intrsect-Chr2^{BNST}* mice, we observed a significant reduction in YFP⁺ cells in the ventral BNST (Extended Data Fig. 8d–f), indicating that a large proportion of VTA-projecting and LH-projecting CRF^{BNST} neurons are located in the ventral BNST. We also found that 5-HT robustly depolarized these *Cr^fCre::Intrsect-Chr2^{BNST}* neurons compared to CRF neurons in general (Fig. 2g–i). Furthermore, we observed light-evoked IPSCs in the BNST of *Cr^fCre::Intrsect-Chr2^{BNST}* mice, confirming local GABA release from these neurons (Extended Data Fig. 8g). These results support the existence of a separate population of local CRF^{BNST} neurons that is excited by 5-HT and increases local GABAergic transmission in the BNST, distinct from a population of CRF^{BNST} neurons that project to and release GABA in the VTA or the LH (Extended Data Fig. 8h–j).

To probe the translational relevance of these BNST microcircuits, we adopted a pharmacological approach using SSRIs. SSRIs represent one of the most widely used classes of drugs for psychiatric disorders. One limitation of SSRIs is that acute administration can lead to negative behavioural states^{1,2}, a finding that is recapitulated in rodent models^{3,16–20}. Importantly, the BNST has been demonstrated to be an anatomical site of action for some of the aversive actions of SSRIs in rodents⁴. This provided the opportunity to test our model that 5-HT

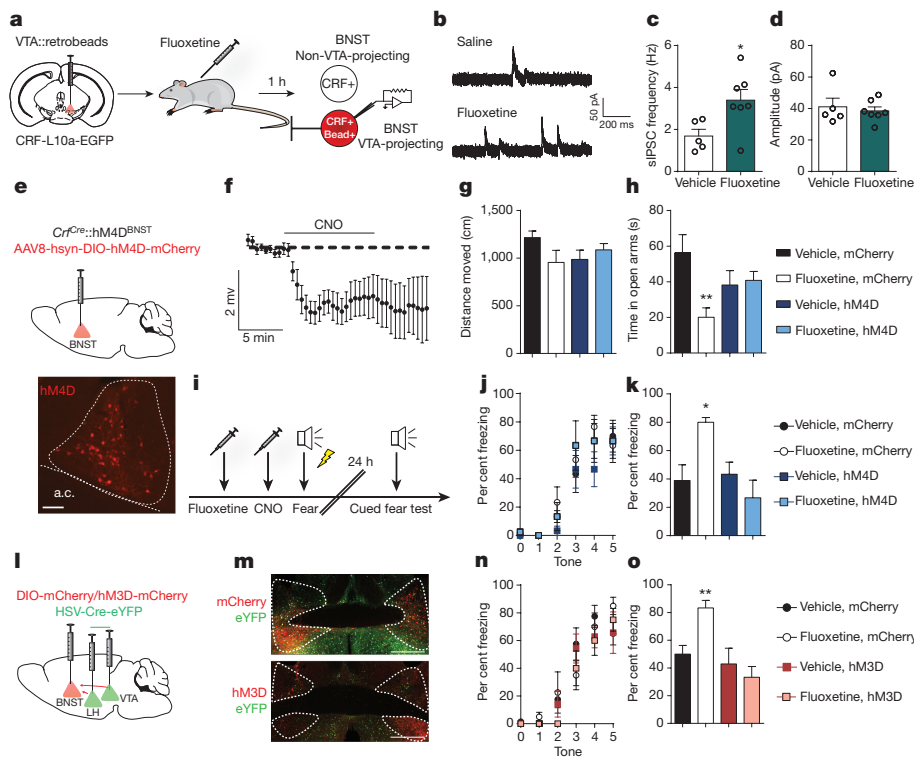


Figure 3 | Acute fluoxetine elicits aversive behaviour by engaging inhibitory CRF circuits in the BNST. **a**, Schematic of recording for *in vivo* fluoxetine experiments in CRF reporter mice. **b**, Representative traces of sIPSCs in VTA-projecting neurons in the BNST for 5 experiments in 2 saline-treated mice and 7 experiments in 2 fluoxetine-treated mice. **c**, **d**, Bar graphs showing that fluoxetine increases in sIPSC frequency ($t_{10} = 2.55$, $P < 0.05$, Student's unpaired two-tailed *t*-test, $n = 5$ cells from 2 saline-treated mice, $n = 7$ cells from 2 fluoxetine-treated mice), but not amplitude ($t_{10} = 0.4752$, $P > 0.05$, Student's unpaired two-tailed *t*-test, $n = 5$ cells from 2 saline mice, $n = 7$ cells from 2 fluoxetine mice) in VTA-projecting neurons in the BNST. **e**, Experimental configuration for assessment of anxiety in fluoxetine-treated *Cr^fCre::hM4Di^{BNST}* (Gi-coupled DREADD) mice and a coronal slice of the BNST expressing hM4Di-mCherry. Scale bar, 100 μ m. **f**, Confirmatory electrophysiology in the BNST showing hyperpolarization of hM4Di-mCherry-expressing cells following bath application of CNO ($t_5 = 4.32$, $P < 0.01$, one-sample *t*-test, $n = 6$ cells from

4 mice). **g**, **h**, Chemogenetic silencing of CRF neurons attenuates fluoxetine-induced anxiety like behaviour on the elevated zero maze ($F_{1,30} = 7.086$, $P < 0.05$, two-way ANOVA, $n = 10$ fluoxetine and hM4Di and $n = 8$ for all other groups) without any concomitant locomotor effects. **i**, Experimental configuration for fear conditioning experiments in *Cr^fCre::hM4Di^{BNST}* mice. **j**, **k**, Chemogenetic silencing of CRF^{BNST} neurons had no effect on freezing behaviour during fear learning but prevented fluoxetine enhancement of cued fear recall ($F_{1,17} = 8.73$, $P < 0.01$, two-way ANOVA, $n = 6$ mCherry and vehicle and $n = 5$ per group for all other groups). **l**, Experimental configuration for assessment of the role of BNST outputs to the VTA and LH in fluoxetine-induced aversive behaviour. **m**, Confocal image of the BNST from HSV^{Cre}::hM3Dq^{BNST} mice. Scale bars, 500 μ m. **n**, **o**, Chemogenetic activation of BNST neurons that project to the midbrain did not impact fear acquisition but attenuated fluoxetine-induced enhancement of cued fear recall ($F_{1,27} = 7.541$, $P < 0.05$, two-way ANOVA, $n = 7$ vehicle/hM3D and $n = 8$ for all other groups). Data are mean \pm s.e.m. * $P < 0.05$; ** $P < 0.01$.

in the BNST drives aversive behaviour through inhibition of BNST outputs to the VTA. We observed that an acute systemic injection of the SSRI fluoxetine increased GABAergic transmission on to VTA projecting neurons in the BNST (Fig. 3a–d). We then interrogated the role of CRF^{BNST} neurons in acute fluoxetine-enhanced anxiety using *Cr^f^{Cre}* mice transduced in the BNST using a Cre-inducible DREADD coupled to the G_o signalling pathway (hM4Di-DREADD). We found that acute fluoxetine potentiated anxiety-like behaviour, and this effect was blocked by chemogenetic inhibition of CRF^{BNST} neurons (Fig. 3e–h).

To evaluate directly whether endogenous 5-HT acts on CRF^{BNST} neurons to enhance cued fear memory, we used the same chemogenetic approach to silence CRF^{BNST} neurons during fluoxetine treatment and subsequent fear conditioning (Fig. 3i). Chemogenetic inhibition of CRF^{BNST} neurons also significantly attenuated fluoxetine-induced enhancement of cued fear recall, providing proof of concept that augmentation of 5-HT via acute SSRI treatment recruits CRF^{BNST} neurons to enhance fear-related behaviour (Fig. 3j, k). Using connectivity based chemogenetic approaches, we then tested whether inhibition of BNST outputs to the VTA and LH is a critical component of 5-HT→BNST-induced aversive states. We observed that activation of G_q signalling in VTA-projecting and LH-projecting BNST neurons, targeted by HSV-Cre-eYFP infused in the VTA and LH and Cre-dependent G_q-coupled DREADD infused in the BNST (HSV^{Cre}::hM3Dq^{BNST}), significantly attenuated fluoxetine enhancement of cued fear recall (Fig. 3l–o). Together, these data provide compelling evidence that acute fluoxetine engenders aversive behaviour by recruiting CRF neurons in the BNST that in turn inhibit putative GABAergic (anxiolytic and stress buffering) outputs from the BNST to the VTA and LH. Pharmacological interventions that target this circuit may improve adverse symptoms during the initial weeks of SSRI treatment. Based on the critical role for CRF^{BNST} neurons in fluoxetine-induced aversive behaviour, we examined the effect of a systemic CRF₁R antagonist on SSRI enhancement of cued fear recall. Blocking the CRF system reduced this aversive state and abolished the increase in sIPSCs in LH-projecting neurons in the BNST during bath application of 5-HT (Extended Data Fig. 9). This provides preclinical evidence that CRF₁R antagonists given in concert with SSRIs could be a promising treatment for anxiety disorders.

Together, these data reveal a discrete 5-HT responsive circuit in the BNST that underlies pathological anxiety and fear associated with a hyperserotonergic state (Extended Data Fig. 10). SSRIs are currently a first-line treatment for anxiety and panic disorders, but can acutely exacerbate symptoms, resulting in poor therapeutic compliance. Our results strongly implicate 5-HT engagement of a local BNST-inhibitory microcircuit in acute SSRI-induced aversive behaviours in rodents, and could potentially be involved in the early adverse events seen in clinical populations, emphasizing the need to identify compounds that selectively target both genetically defined and pathway-specific cell populations.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 February 2015; accepted 20 July 2016.

Published online 24 August 2016.

- Gorman, J. M. *et al.* An open trial of fluoxetine in the treatment of panic attacks. *J. Clin. Psychopharmacol.* **7**, 329–332 (1987).
- Westenberg, H. G. M. & den Boer, J. A. Serotonin-influencing drugs in the treatment of panic disorder. *Psychopathology* **22** (Suppl 1), 68–77 (1989).
- Burghardt, N. S., Bush, D. E., McEwen, B. S. & LeDoux, J. E. Acute selective serotonin reuptake inhibitors increase conditioned fear expression: blockade with a 5-HT_{2c} receptor antagonist. *Biol. Psychiatry* **62**, 1111–1118 (2007).
- Ravinder, S., Burghardt, N. S., Brodsky, R., Bauer, E. P. & Chattarji, S. A role for the extended amygdala in the fear-enhancing effects of acute selective serotonin reuptake inhibitor treatment. *Transl. Psychiatry* **3**, e209 (2013).
- Phelix, C. F., Liposits, Z. & Paul, W. K. Serotonin–CRF interaction in the bed nucleus of the stria terminalis: a light microscopic double-label immunocytochemical analysis. *Brain Res. Bull.* **28**, 943–948 (1992).
- Liu, Z. *et al.* Dorsal raphe neurons signal reward through 5-HT and glutamate. *Neuron* **81**, 1360–1374 (2014).

- Burke, L. K. *et al.* Sex difference in physical activity, energy expenditure and obesity driven by a subpopulation of hypothalamic POMC neurons. *Mol. Metab.* **5**, 245–252 (2016).
- Armbruster, B. N., Li, X., Pausch, M. H., Herlitze, S. & Roth, B. L. Evolving the lock to fit the key to create a family of G protein-coupled receptors potentially activated by an inert ligand. *Proc. Natl Acad. Sci. USA* **104**, 5163–5168 (2007).
- Heisler, L. K., Zhou, L., Bajwa, P., Hsu, J. & Tecott, L. H. Serotonin 5-HT_{2c} receptors regulate anxiety-like behavior. *Genes Brain Behav.* **6**, 491–496 (2007).
- Olive, M. F., Koenig, H. N., Nannini, M. A. & Hodge, C. W. Elevated extracellular CRF levels in the bed nucleus of the stria terminalis during ethanol withdrawal and reduction by subsequent ethanol intake. *Pharmacol. Biochem. Behav.* **72**, 213–220 (2002).
- Huang, M. M. *et al.* Corticotropin-releasing factor (CRF) sensitization of ethanol withdrawal-induced anxiety-like behavior is brain site specific and mediated by CRF-1 receptors: relation to stress-induced sensitization. *J. Pharmacol. Exp. Ther.* **332**, 298–307 (2010).
- Jennings, J. H. *et al.* Distinct extended amygdala circuits for divergent motivational states. *Nature* **496**, 224–228 (2013).
- Dabrowska, J. *et al.* Neuroanatomical evidence for reciprocal regulation of the corticotropin-releasing factor and oxytocin systems in the hypothalamus and the bed nucleus of the stria terminalis of the rat: implications for balancing stress and affect. *Psychoneuroendocrinology* **36**, 1312–1326 (2011).
- Kim, S.-Y. *et al.* Diverging neural pathways assemble a behavioural state from separable features in anxiety. *Nature* **496**, 219–223 (2013).
- Fenno, L. E. *et al.* Targeting cells with single vectors using multiple-feature Boolean logic. *Nature Methods* **11**, 763–772 (2014).
- Dekeyne, A., Denorme, B., Monneyron, S. & Millan, M. J. Citalopram reduces social interaction in rats by activation of serotonin (5-HT)_{2c} receptors. *Neuropharmacology* **39**, 1114–1117 (2000).
- Belzung, C., Le Guisquet, A. M., Barreau, S. & Calatayud, F. An investigation of the mechanisms responsible for acute fluoxetine-induced anxiogenic-like effects in mice. *Behav. Pharmacol.* **12**, 151–162 (2001).
- Javelot, H. *et al.* Efficacy of chronic antidepressant treatments in a new model of extreme anxiety in rats. *Depress. Res. Treat.* **2011**, 531435 (2011).
- Liu, J. *et al.* Acute administration of leptin produces anxiolytic-like effects: a comparison with fluoxetine. *Psychopharmacology* **207**, 535–545 (2010).
- Mombereau, C., Gur, T. L., Onksen, J. & Blendy, J. A. Differential effects of acute and repeated citalopram in mouse models of anxiety and depression. *Int. J. Neuropsychopharmacol.* **13**, 321–334 (2010).

Acknowledgements We acknowledge B. Roth for providing DREADD viral constructs and *Sert^{Cre}* mice, and B. Lowell for providing *Cr^f^{Cre}* mice. We also thank A. Lopez, D. Perron, and A. Kendra for technical assistance with stereotaxic surgeries on mice, B. Geenen for technical assistance with immunohistochemistry and E. Dankoski for technical assistance with the FSCV. This work was supported by NIH grants AA019454, AA011605 (T.L.K.), the Wellcome Trust (098012) and the Biotechnology and Biological Sciences Research Council grant (BB/K001418/1) (L.K.H.) and by NIH grant K01AA023555 and the Alcohol Beverage Medical Research Fund (Z.A.M.). C.A.M. was supported by a postdoctoral NIAAA F32 fellowship (AA021319-02). C.M.M. is supported by a predoctoral NIAAA F31 fellowship (F31AA023440).

Author Contributions C.A.M., C.M.M., G.D., Z.A.M., L.K.H. and T.L.K. designed the experiments. A.H. and J.F.D. performed triple label fos/tryptophan hydroxylase/Fluor-Gold staining and image analysis. L.R.H. performed electrode placement surgeries and *in vivo* recordings during fear acquisition and recall. C.A.M. performed stereotaxic surgeries for evoked 5-HT electrophysiology and optogenetic behavioural experiments. Z.A.M. performed slice FSCV experiments and C.A.M. performed evoked 5-HT electrophysiology experiments. C.A.M. performed stereotaxic surgeries, behavioural and data analysis for 5-HT^{DRN}→BNST optogenetic experiments. C.A.M. performed all slice electrophysiology experiments and C.M.M. and C.A.M. performed stereotaxic surgeries for these experiments (retrograde tracers, ChR2 infusions, and hM3D and hM4D infusions). C.M.M. performed stereotaxic surgeries for chemogenetic manipulations in CRF^{BNST} neurons that were used in fluoxetine fear conditioning experiments and C.A.M. performed behavioural and data analysis. C.E.D. performed surgeries for electrophysiological recordings and data analysis for fear conditioning experiments. M.N. and J.F.D. performed surgeries for chemogenetic manipulations in CRF^{BNST} neurons that were used in fluoxetine anxiety (EZM) assays and N.B. and C.A.M. performed behavioural and data analysis. C.M.M. and J.F.D. performed stereotaxic surgeries for HSV^{Cre}::hM3D^{BNST} behavioural manipulations and C.A.M. performed behavioural and data analysis. C.M.M. also performed imaging and analysis for optogenetic experiments, chemogenetic, and Intrex experiments. C.R. and K.D. designed Intrex viral constructs. G.D. and C.C. performed surgeries, behavioural and data analysis for *Htr2c^{Cre}*::hM3D^{BNST} experiments. C.A.M., C.M.M. and T.L.K. wrote the manuscript with input from Z.A.M., L.R.H., J.F.D., J.A.H., G.D., T.E.T., A.H., L.K.H. and T.K.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.L.K. (thomas_kash@med.unc.edu).

Reviewer Information *Nature* thanks A. Sahay and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Data reporting. Based on power analyses that assumed a normal distribution, a 20% change in mean and 15% variation, we determined that at least 9 mice per group would be needed for behavioral experiments. This was adhered to as far as possible, except in cases where mice had to be removed owing to misplaced injections or lost headcaps. Mice were randomly assigned to groups and attempts were made to balance groups according to variables such as age and housing condition. The investigators were not blinded to allocation during experiments, but were blinded to outcome assessment for all behavioral experiments.

Mice. Mice were used in all experiments. For experiments involving Cre lines, mice were crossed for several generations to C57 mice before using. All wild-type mice were C57BL/6 mice obtained from The Jackson Laboratory (Bar Harbour, ME). For all behavioural experiments except those involving *Htr2c^{Cre}* mice, male mice ranging in age from 8–16 weeks were used. Female *Htr2c^{Cre}* mice were used in chemogenetic manipulations. Both male and female mice aged 6–20 weeks were used for slice electrophysiology and anatomical tracing experiments. All behavioural studies or tissue collection for *ex vivo* slice electrophysiology were performed during the light cycle.

All behavioural experiments in *Htr2c^{Cre}* mice were conducted at the University of Aberdeen and in accordance with the United Kingdom Animals (Scientific Procedures) Act of 1986. All *in vivo* electrophysiology experiments were conducted in accordance with all rules and regulations at the National Institute for Alcohol Abuse and Alcoholism at the National Institutes of Health. All other procedures were conducted in accordance with the National Institutes of Health guidelines for animal research and with the approval of the Institutional Animal Care and Use Committee at the University of North Carolina at Chapel Hill.

All animals were group housed on a 12 h light cycle (lights on at 7 a.m.) with *ad libitum* access to rodent chow and water, unless described otherwise. CRF-ires-Cre (*Cr^fCre*) were provided by Bradford Lowell (Harvard University) and were previously described²¹. C57BL/6J mice were obtained from the Jackson Laboratory (Bar Harbour, ME). To visualize CRF-expressing neurons, *Cr^fCre* mice were crossed with either an Ai9 or a Cre-inducible L10-GFP reporter line (Jackson Laboratory)²² to produce CRF-Ai9 or CRF-L10GFP progeny, referred to throughout the manuscript as CRF-reporters. *Sert^{Cre}* mice (from GENSAT) were a generous gift from Bryan Roth. *Htr2c^{Cre}* mice were supplied by Lora Heisler and are described in detail elsewhere⁷.

Male mice were used for *in vivo* optogenetic behavioural experiments and for assessing the involvement of BNST CRF neurons on fluoxetine-induced enhancement of fear. Female 5-HT_{2C}-Cre mice were used in chemogenetic manipulations. Both male and female mice were used for slice electrophysiology and anatomical tracing experiments. All behavioural studies or tissue collection for *ex vivo* slice electrophysiology were performed during the light cycle.

Viruses and tracers. All AAV viruses except INTRSECT constructs were produced by the Gene Therapy Center Vector Core at the University of North Carolina at Chapel Hill and had titres of >10¹² genome copies per ml. For *ex vivo* and *in vivo* optical experiments, mice were injected with rAAV5-eYFP or rAAV5-eYFP as a control. Red IX retrobeads (Lumafluor) were used to fluorescently label LH- and VTA-projecting BNST neurons during *ex vivo* slice electrophysiology recordings. The retrograde tracer Fluoro-Gold (Fluorochrome) was used for anatomical mapping. Cholera toxin B (CTB) 555 and CTB 657 retrograde tracers (Invitrogen; C34776, and C34778, respectively) diluted to 0.5% (w/v) in sterile PBS were used per injection site for anatomical mapping of collateral projections from BNST to LH and VTA. For chemogenetic manipulations, mice were injected with 400 nl of rAAV8-hsyn-DIO-hM3D(Gq)-mCherry, rAAV8-hsyn-DIO-hM4D(Gi)-mCherry, or rAAV8-hsyn-DIO-mCherry bilaterally. HSV-hEF1 α -mCherry, HSV-eYFP-LSL1-mCherry-IRES-flo, and HSV-eYFP-IRES-Cre (supplied by Rachel Neve at the McGovern Institute for Brain Research at MIT) were injected bilaterally into the VTA and LH at a volume of 500 nl per site. The INTRSECT construct AAVdj-hSyn-Con/Foff-hChR2(H134R)-EYFP was infused at 500 nl per side into the BNST. All AAV constructs had viral titres >10¹² genome particles per ml.

Stereotaxic injections. All surgeries were conducted using aseptic technique. Adult mice (2–5 months) were deeply anaesthetized with 5% isoflurane (v/v) in oxygen and placed into a stereotaxic frame (Kopf Instruments) while on a heated pad. Sedation was maintained at 1.5–2.5% isoflurane during surgery. An incision was made down the midline of the scalp and a craniotomy was performed above the target regions and viruses and fluorescent tracers were microinjected using a Neuros Hamilton syringe at a rate of 100 nl min⁻¹. After infusion, the needle was left in place for 10 min to allow for diffusion of the virus before the needle was slowly withdrawn. Injection coordinates (in mm, midline, Bregma, dorsal surface): BNST (± 1.00 , 0.30, -4.35), LH (± 0.9 to 1.10, -1.7 , -5.00 to -5.2), VTA (-0.3 , -2.9 , -4.6), DR (0.0, -4.65 , -3.2 with a 23° angle of approach). When using retrobeads, injection volumes into the LH and VTA were 300 nl and

400 nl, respectively. Fluoro-Gold injection volumes were 200 nl per target site. CTB volumes were 200 nl per target site. An optical fibre was implanted in the BNST (± 1.00 , 0.20, -4.15) at a 10° angle for *in vivo* photostimulation studies. After fibre implantation, dental cement was used to adhere the ferrule to the skull. Following surgery, all mice returned to group housing. Mice were allowed to recover for at least 3 weeks before being used for chemogenetic behavioural studies, or 6 weeks for *in vivo* optogenetic studies.

Drugs. RS-102221, 5-HT and mCPP were from Tocris (Bristol, UK). For electrophysiology experiments, RS-102221 was made up to 100 mM in DMSO and then diluted to a final concentration of 5 μ M in aCSF. 5-HT and mCPP were stocked at 10 and 20 mM, respectively, in ddH₂O and diluted to their final concentrations in aCSF. For electrophysiology experiments, clozapine-N-oxide (CNO; from Bryan Roth) was stocked at 100 mM in DMSO and diluted to 10 μ M in aCSF. For behaviour experiments, CNO was dissolved in 0.5% DMSO (in 0.9% saline) to a concentration of 0.1 mg ml⁻¹ or 0.3 mg ml⁻¹ and injected at 10 ml per kg for a final concentration of 1 or 3 mg per kg, i.p. Fluoxetine (Sigma) was made up in 0.9% NaCl to a concentration of 1 mg ml⁻¹ and then injected at 10 ml per kg for a final concentration of 10 mg per kg, i.p.

In vivo electrophysiological procedures. Surgical procedures. Mice were anaesthetized with 2% isoflurane (Baxter Healthcare, Deerfield, IL) and implanted with 2 \times 8 electrode (35 μ m tungsten) micro-arrays (Innovative Neurophysiology, Durham, NC) targeted at the BNST (ML: 0.8 mm, AP: ± 0.5 mm, and DV: -4.15 mm relative to Bregma). Following surgery, mice were singly housed and allowed at least one week to recover before behavioural testing.

Fear conditioning. Fear conditioning took place in 27 \times 27 \times 11 cm conditioning chambers (Med Associates, St. Albans, VT), with a metal-rod floor (context A) and scented with 1% vanilla. Mice received 5 pairings of a pure tone CS with a 0.6 mA foot shock. 24 h following conditioning, mice underwent a CS recall test (10 presentations of the CS alone, 5 s ITI), which was conducted in a Plexiglas cylinder (20 cm diameter) and scented with 1% acetic acid (context B). Stimulus presentations for both tests were controlled by MedPC (Med Associates, St. Albans, VT). Cameras were mounted overhead for recording freezing behaviour, which was scored automatically using CinePlex Behavioural Research System software (Plexon, Dallas, TX).

Electrophysiological recording and single unit analysis. Electrophysiological recording took place during both fear conditioning and CS recall tests. Individual units were identified and recorded using Omniplex Neural Data Acquisition System (Plexon, Dallas, TX). Neural data was sorted using Offline Sorter (Plexon, Dallas, TX). Waveforms were isolated manually, using principal component analysis. To be included in the analyses, spikes had to exhibit a refractory period of at least 1 ms. Autocorrelograms from simultaneously recorded units were examined to ensure that no cell was counted twice. Single units were analysed by generating perievent histograms (3 s bins) of firing rates from 30 s before CS onset until 30 s after CS offset (NeuroExplorer 5.0, Nex Technologies, Madison, AL). Firing rates were normalized to baseline (30 s before CS onset) using z-score transformation. Analysis included a total of 139 cells over three days of recording. Data reported for raw firing rates include only putative principal neurons (<10 Hz).

The formula for computing the suppression ratio was (average freezing rate) / (average freezing rate + average movement rate). Each cell was calculated individually. A value of 0.5 = no change in rate).

Ex vivo slice electrophysiology. Brains were sectioned at 0.07 (mm per s) on a Leica 1200S vibratome to obtain 300 μ m coronal slices of the BNST, which were incubated in a heated holding chamber containing normal, oxygenated aCSF (in mM: 124 NaCl, 4.4 KCl, 2 CaCl₂, 1.2 MgSO₄, 1 NaH₂PO₄, 10.0 glucose, and 26.0 NaHCO₃) maintained at 30 \pm 1 °C for at least 1 h before recording. Slices were transferred to a recording chamber (Warner Instruments) submerged in normal, oxygenated aCSF maintained at 28–30 °C at a flow rate of 2 ml min⁻¹. Neurons of the BNST were visualized using infrared differential interference contrast (DIC) video-enhanced microscopy (Olympus). Borosilicate electrodes were pulled with a Flaming-Brown micropipette puller (Sutter Instruments) and had a pipette resistance between 3–6 M Ω . Signals were acquired via a Multiclamp 700B amplifier, digitized at 10 kHz and analysed with Clampfit 10.3 software (Molecular Devices, Sunnyvale, CA, USA).

Light-evoked action potentials. In *Sert^{Cre}* or *Cr^fCre* mice, fluorescently labelled neurons expressing ChR2 were visualized and stimulated with a blue (470 nm) LED using a 1 Hz, 2 Hz, 5 Hz, 10 Hz, and 20 Hz stimulation protocol with a pulse width of 0.5 ms. Evoked action potentials were recorded in current clamp mode using a potassium gluconate based internal solution (in mM: 135 K⁺ gluconate, 5 NaCl, 2 MgCl₂, 10 HEPES, 0.6 EGTA, 4 Na₂ATP, 0.4 Na₂GTP, pH 7.3, 285–290 mOsmol). **Light-evoked synaptic transmission.** In *Cr^fCre* mice with ChR2 in the BNST and retrograde tracer beads in the VTA or LH, we visualized non-ChR2-expressing, beaded neurons using green (532 nm) LED. Recordings were conducted in voltage

clamp mode using a caesium-methanesulfonate (Cs-Meth) based internal solution (in mM: 135 caesium methanesulfonate, 10 KCl, 1 MgCl₂, 0.2 EGTA, 2 QX-314, 4 MgATP, 0.3 GTP, 20 phosphocreatine, pH 7.3, 285–290 mOsmol) so that we could detect EPSCs (–55 mV) and IPSCs (+10 mV) in the same neuron. After confirming the absence of a light-evoked EPSC signal, we measured light-evoked IPSCs during a single 5-ms light pulse of 470 nm. In a subset of these experiments, SR95531 (GABAzine, 10 μ M) was bath applied for 10 min to block IPSCs.

Drug effects in CRF^{BNST} neurons. Crf-reporter mice were injected with retrograde tracer beads into the VTA (ML –0.5, AP –2.9, DV –4.6). We then recorded from beaded (VTA-projecting) and non-beaded (non-projecting) CRF neurons in the BNST. Acute drug effects were determined in current clamp mode in the presence of TTX using a potassium gluconate-based internal solution. After a 5-min stable baseline was established, 5-HT (10 μ M) or mCPP (20 μ M) was bath applied for 10 min while recording changes in membrane potential. The difference in membrane potential between baseline and drug application at peak effect (Δ MP) was later determined. In a subset of mCPP experiments, slices were incubated with RS-102221 (5 μ M) for at least 20 min before experiments began.

Synaptic transmission. Spontaneous inhibitory postsynaptic currents (sIPSCs) were assessed in voltage clamp using a potassium-chloride gluconate-based intracellular solution (in mM: 70 KCl, 65 K⁺-gluconate, 5 NaCl, 10 HEPES, 0.5 EGTA, 4 ATP, 0.4 GTP, pH 7.2, 285–290 mOsmol). IPSCs were pharmacologically isolated by adding kynurenic acid (3 mM) to the aCSF to block AMPA and NMDA receptor-dependent postsynaptic currents. The amplitude and frequency of sIPSCs were determined from 2 min recording episodes at –70 mV. The baseline was averaged from the 4 min preceding the application of 5-HT (10 μ M) or mCPP (10 μ M) for 10 min. In a subset of these experiments, RS-102221 (5 μ M) was added to the aCSF and slices were incubated in this drug solution for at least 20 min before experiments began. For miniature IPSCs (mIPSCs), TTX was included in the aCSF to block network activity.

In *Sert^{Cre};*ChR2^{BNST} mice with retrograde tracer beads in the VTA, sIPSCs were recorded as described above. After achieving a stable baseline, a 10 s, 20 Hz photostimulation was applied.

For assessment of spontaneous excitatory postsynaptic currents (sEPSCs), a caesium gluconate-based intracellular solution was used (in mM: 135 Cs⁺-gluconate, 5 NaCl, 10 HEPES, 0.6 EGTA, 4 ATP, 0.4 GTP, pH 7.2, 290–295 mOsmol). AMPA_R-mediated EPSCs were pharmacologically isolated by adding 25 μ M picrotoxin to the aCSF. sEPSC recordings were acquired in 2 min recording blocks at –70 mV.

Fast-scan cyclic voltammetry (FSCV). Electrodes were fabricated as previously described and cut to 50–100 μ m in length²³. Animal and slice preparation were as described above for electrophysiology and slices were perfused on the rig in ACSF. Using a custom-built potentiostat (University of Washington Seattle), 5-HT recordings were made in the BNST using TarHeel CV written in laboratory view (National Instruments). Briefly a triangular waveform (–0.1 V to 1.3 V with a 10% phase shift at 1,000 V per s, versus Ag/AgCl) was applied to the carbon fibre electrode at a rate of 10 Hz. Slices were optically stimulated with 20 5-ms blue (490 nm) light pulses at a rate of 20 Hz down the submerged 40 \times objective. 10 cyclic voltammograms were averaged before optical stimulation for background subtraction. Voltammograms were digitally smoothed one time with a fast Fourier transform following data collection and analysed with HDCV (UNC Chapel Hill). Fluoxetine (10 μ M) was bath applied following a stable baseline (20 min).

Behavioural assays. For chemogenetic manipulations, mice were transported to a holding cabinet adjacent to the behavioural testing room to habituate for at least 30 min before being pretreated with CNO (3 mg per kg, i.p. for *Crf^{Cre}* mice and 1 mg per kg, i.p. for *Htr2C^{Cre}* mice). All behavioural testing began 45 min following CNO treatment, with the exception of fear conditioning training, which occurred 30 min after a CNO injection. When assessing the effect of fluoxetine on fear conditioning, fluoxetine (10 mg per kg, i.p.), or vehicle, was administered 1 h before training (30 min before CNO treatment). For optogenetic manipulations, mice received bilateral stimulation (473 nm, ~10 mW, 5 ms pulses, 20 Hz) when specified. Unless specified, all equipment was cleaned with a damp cloth between mouse trials. All sessions were video recorded and analysed using EthoVision software (Noldus Information Technologies) except where noted.

Elevated plus maze. Mice were placed in the centre of an elevated plus maze and allowed to explore during a 5 min session. Light levels in the open arms were ~14 lux. During optogenetic manipulations mice received bilateral stimulation during the entire 5 min session. Mice that left the maze were excluded from analysis ($n = 2$ control, 1 ChR2 from optogenetic experiments).

Open field. Mice were placed into the corner of a white Plexiglas open field arena (25 \times 25 \times 25 cm) and allowed to freely explore for 30 min. The centre of the open field was defined as the central 25% of the arena. For optogenetic studies the 30 min session was divided into three 10-min epochs consisting of stimulation off, stimulation on, and stimulation off periods.

Novelty-induced suppression of feeding. 48 h before testing, mice were provided with access to a single piece of Froot Loops cereal (Kellogg's) in their home cage. 24 h before testing, home cage chow was removed and mouse body weights were recorded. Water remained available *ad libitum*. Beginning at least one hour before testing, mice transferred to new clean cages so they were singly housed for the test session and body weights were recorded. During the test session mice were placed into an arena (25 \times 25 \times 25 cm) that contained a single Froot Loop on top of a piece of circular filter paper. Mice were monitored by a live observer and the latency for the mouse to begin eating the pellet was measured, allowing up to 10 min. All mice began eating within this time. Following the initiation of feeding, mice were removed from the arena and placed back into their home cages. Mice were then provided with 10 min of access to a pre-weighed amount of Froot Loops for a post-test feeding session. After this 10 min post-test, the remaining Froot Loops were weighed and mice were returned to *ad libitum* home cage chow. Mice were returned to group housing at the end of this session. For optogenetic experiments, mice received constant 20 Hz optical stimulation during both the latency to feed assay and the 10 min post-test. During optogenetic experiments, one control mouse did not feed during the 10 min NSF session and was excluded from the results.

Home cage feeding. *Sert^{Cre}* mice were food deprived for 24 h. On the day of the experiment, mice were acclimated to the behaviour room for 1 h. A single pre-weighed food pellet was placed in the home cage and the mice were allowed to eat for 10 min during optogenetic stimulation. At the end of the experimental session, the pellet was removed and weighed and mice were given *ad libitum* access to food.

Htr2C^{Cre} mice were acclimated in metabolic chambers (TSE Systems, Germany) for 2 days before the start of the recordings. After acclimation, mice were food deprived for 24 h. Following fasting, mice received an i.p. injection of CNO 30 min before food presented again. Mice were recorded for 12 h with the following measurements being taken every 30 min: water intake, food intake, ambulatory activity (in x and z axes), and gas exchange (O₂ and CO₂) (using the TSE LabMaster system, Germany). Energy expenditure was calculated according to the manufacturer's guidelines (PhenoMaster Software, TSE Systems).

Fear conditioning. We used a three-day protocol to assess both cued and contextual fear recall. On the first day, mice were placed into a fear conditioning chamber (Med Associates) that contained a grid floor and was cleaned with a scented paper towel (19.5% ethanol, 79.5% H₂O, 1% vanilla). After a 3 min baseline period, mice were exposed to a 30 s tone (3 kHz, 80 dB) that co-terminated with a 2 s scrambled foot shock (0.6 mA). A total of 5 tone-shock pairings were delivered with a random inter-tone interval (ITI) of 60–120 s. For optogenetic studies, light stimulation occurred only during the 30-s tones of this session. Following delivery of the last foot shock, mice remained in the conditioning chamber for a 2-min consolidation period. 24 h later, mice were placed into a separate conditioning box (Med Associates) that contained a white Plexiglas floor, a striped pattern on the walls, and was cleaned and scented with a 70% ethanol solution. After a 3 min baseline period, mice were presented with 10 tones (30 s, 3 kHz, 80 dB) with a 5 s ITI. Mice remained in the chamber after the last tone for a two-minute consolidation period. 24 h later (48 h after training), mice were returned to the original training chamber for 5 min. For each session, freezing behaviour was hand-scored every 5 s by a trained observer blinded to experimental treatment as described previously²⁴. Freezing was defined as a lack of movement except as required for respiration.

Immunohistochemistry and histology. All mice used for behavioural and anatomical tracing experiments were anesthetized with Avertin and transcardially perfused with 30 ml of ice-cold 0.01 M PBS followed by 30 ml of ice-cold 4% paraformaldehyde (PFA) in PBS. Brains were extracted and stored in 4% PFA for 24 h at 4 °C before being rinsed twice with PBS and stored in 30% sucrose and PBS until the brains sank. 45 μ m slices were obtained on a Leica VT100S and stored in 50/50 PBS/Glycerol at –20 °C. DREADD or ChR2-containing sections were mounted on slides, allowed to dry, coverslipped with VectaShield (Vector Labs, Burlingame, CA), and stored in the dark at 4 °C.

Tryptophan hydroxylase/Fluoro-Gold/c-fos triple labelling. We stained free-floating dorsal raphe sections using indirect immunofluorescence sequentially for first tryptophan hydroxylase (TPH) and Fluoro-Gold (FG) and then *c-fos*. For TPH/FG, we washed sections 3 \times for 5 min with 0.01 M PBS, permeabilized them for 30 min in 0.5% Triton/0.01 M PBS, and washed the sections again 2 \times with 0.01 M PBS. We blocked the sections for 1 h in 0.1% Triton/0.01 M PBS containing 10% (v/v) normal donkey serum and 1% (w/v) bovine serum albumin (BSA). We then added primary antibodies (1:500 mouse anti-TPH (Sigma Aldrich T0678) and 1:3,000 guinea-pig anti-Fluoro-Gold (Protos Biotech NM101)) to blocking buffer and incubated the sections overnight at 4 °C. The next day, we washed the sections 3 \times for 5 min with 0.01 M PBS, then incubated them with 1:500 with Alexa Fluor 647-conjugated donkey anti-mouse and Alexa Fluor 488-conjugated donkey anti-guinea pig secondary antibodies for 2 h at room temperature, and washed the sections 4 \times for 5 min with 0.01 M PBS. We then proceeded directly to the *c-fos* tyramide signal amplification based immunofluorescent staining. We permeabilized

the sections in 50% methanol for 30 min, then quenched endogenous peroxidase activity in 3% hydrogen peroxide for 5 min. Followed by two 10 min washes in 0.01 M PBS, we blocked the sections in PBS containing 0.3% Triton X-100 and 1.0% BSA for 1 h. c-fos primary antibody (Santa Cruz Biotechnology, sc-52) was added to sections at 1:3,000 and sections were incubated for 48 h at 4°C. On day 3, we washed the sections in TNT buffer (0.1 M Tris-HCl pH 7.5, 0.15 M NaCl, 0.05% Tween-20) for 10 min, blocked in TNB buffer (0.1 M Tris-HCl pH 7.5, 0.15 M NaCl, 0.5% blocking reagent – PerkinElmer FP1020) buffer for 30 min. We then incubated the sections in secondary antibody (goat anti-rabbit HRP-conjugated PerkinElmer) 1:200 in TNB buffer for 30 min, washed the sections in TNT buffer 4× for 5 min, and then incubated the sections in Cy3 dye diluted in TSA amplification diluents for 10 min. We washed the sections 2× in TNT buffer, mounted them on microscope slides. We coverslipped the slides using Vectashield mounting medium. We acquired 4–5 of 2 × 4 tiled z-stack (5 optical slices comprising 7 µm total) images of the dorsal raphe from each naive and shock mouse on a Zeiss 800 upright confocal microscope. Scanning parameters and laser power were matched between groups. Images were preprocessed using stitching and maximum intensity projection and then analysed using an advanced processing module in Zeiss Zen Blue that allows nested analysis of multiple segmented fluorescent channels within parent classes. Double-labelled and triple-labelled cells were validated in a semi-automated fashion. At least 4 sections per mouse were counted in this way. One mouse was identified as a significant outlier in the shock group and was excluded from further analysis.

Sert^{Cre}::ChR2, and Crf^{Intsect}-ChR2 validation. To verify expression of ChR2-expressing fibres in the BNST originating from DRN serotonergic neurons, 300 µm slices used for *ex vivo* electrophysiological recordings containing the DRN and BNST were stored in 4% paraformaldehyde at 4°C for 24 h before being rinsed with PBS, mounted, and coverslipped with Vectashield mounting medium. Images showing eYFP fluorescence from the DRN and BNST were obtained on a Zeiss 800 upright confocal microscope using a 10× objective and tiled z stacks. To validate the INTRSECT construct, mice received injections of HSV-hEF1α-mCherry or HSV-ef1α-LSL1-mCherry-IRES-flpo to both the LH and VTA bilaterally ($n = 4$ and 5, respectively). Both groups received AAVDJ-hSyn-Cre-on/Flp-off-hChR2(H134R)-EYFP to the BNST bilaterally. Six weeks following injection, mice were perfused and tissue was collected as described above. To visualize YFP expression in the BNST of Crf^{Cre}::Intsect^{BNST} mice, free-floating slices containing the BNST were rinsed three times with PBS for 5 min each. Slices were then incubated in 50% methanol for 30 min then incubated in 3% hydrogen peroxide for 5 min. Following three 10-min washes in PBS, slices were incubated in 0.5% Triton X-100 for 30 min followed by a 10 min PBS wash. Slices were blocked in 10% normal donkey serum/0.1% Triton X-100 for 1 h, and then they were incubated overnight at 4°C with a primary chicken anti-GFP antibody (GFP-1020, Aves) at 1:500 in blocking solution. Following primary incubation, slices were rinsed three times with 0.01M PBS for 10 min each and incubated with a fluorescent secondary antibody (AlexaFluor 488 donkey anti-chicken) at 1:200 in PBS for 2 h at room temperature. Slices were then rinsed with four 10-min PBS washes before being mounted onto glass slides and coverslipped with Vectashield with DAPI. A 3 × 4 tiled z stack (7 optical sections comprising 35 µm total) image from both the left and right hemispheres of the BNST was obtained at 20× magnification using a Zeiss 800 upright confocal microscope. Scanning parameters and laser power were matched between groups. Images were preprocessed using stitching and maximum-intensity projection. The number of fluorescent cells in the dorsal and ventral aspects of the BNST were counted by a blinded scorer using the cell counter plug-in in FIJI (ImageJ). Each hemisphere was considered independently per mouse. One mouse in the flp-expressing group was a significant outlier for number of cells expressed in a ventral BNST hemisphere (ROUT, $Q = 0.1\%$) and all data from that mouse were excluded.

Cholera toxin retrograde tracer studies in CRF reporter mice. 3 male CRF-L10a reporter mice were injected with 200 nl of CTB 555 and CTB 647 bilaterally to the LH and VTA, respectively, as described above. 5 days following injection, mice were perfused as described above, the brains were extracted, and were stored in 4% paraformaldehyde for 24 h at 4°C before being rinsed with PBS and transferred to 30% sucrose until the brains sank. 45 µm sections containing the BNST were collected as described above. Sections containing the BNST were mounted on glass slides and coverslipped using Vectashield. An image from the left and right hemispheres of a medial section of the BNST was obtained on a Zeiss 800 upright microscope using a 20× objective and 3 × 5 tiled z stacks (5 optical slices comprising 7 µm total). Images were preprocessed using stitching and maximum intensity projection, and

were then analysed using the cell counter function in FIJI (ImageJ). Only cells positive for GFP (putative CRF neurons) were considered. Cells were scored exclusively as either 555+ only (LH-projecting), 647+ only (VTA-projecting), 555+ and 647+ (projecting to both LH and VTA), or 555– and 647– (unlabelled; neither LH- nor VTA-projecting). The total number of CRF neurons scored was calculated as the sum of all four groups, and percentages of each type were calculated from this value. Each hemisphere was scored and plotted independently ($n = 6$ images from 3 mice), and the dorsal and ventral BNST were considered separately. The average values were plotted as pie charts (Extended Data Fig. 5).

Double fluorescence *in situ* hybridization (FISH). For validation of 2C-cre line and comparison of CRF/2C mRNA cellular co-localization, mice were anesthetized using isoflurane, rapidly decapitated, and brains rapidly extracted. Immediately after removal, the brains were placed on a square of aluminium foil on dry ice to freeze. Brains were then placed in a –80°C freezer for no more than 1 week before slicing. 12 µm slices were made of the BNST on a Leica CM3050S cryostat (Germany) and placed directly on coverslips. FISH was performed using the Affymetrix ViewRNA 2-Plex Tissue Assay Kit with custom probes for CRF, 5-HT_{2C}, and Cre designed by Affymetrix (Santa Clara, CA). Slides were coverslipped with SouthernBiotech DAPI Fluoromount-G. (Birmingham, AL). 3 × 5 tiled z stack (15 optical sections comprising 14 µm total) images of the entire 12 µm slice were obtained on a Zeiss 780 confocal microscope for assessment of CRF/2C colocalization. A single-plane 40× tiled image of a CRF/2C slice was obtained on a Zeiss 800 upright confocal microscope for the magnified image shown in Extended Data 6b, right. 3 × 5 tiled z stack (7 optical sections comprising 18 µm) images of 2C/Cre slices were obtained on a Zeiss 800 upright confocal microscope for the 2C/Cre validation. All images were preprocessed with stitching and maximum intensity projection. An image of the BNST from 3 mice in each condition was hand counted for each study using the cell counter plugin in FIJI (ImageJ). Cells were classified into three groups: probe 1+, probe 2+, or probe 1 and 2+. Only cells positive for a probe were considered. Results are plotted as average classified percentages across the three images.

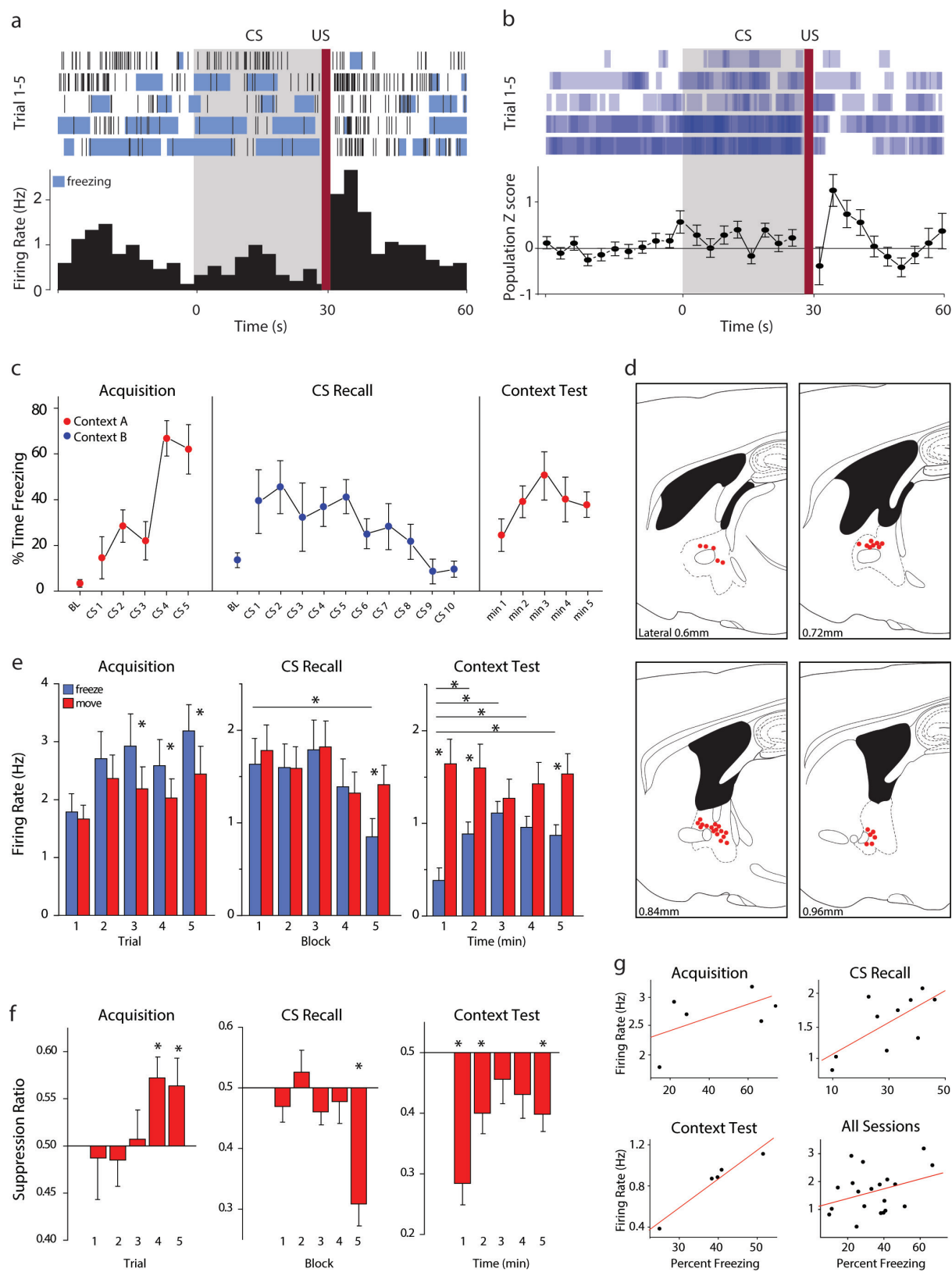
Group assignment. No specific method of randomization was used to assign groups. Animals were assigned to experimental groups so as to minimize the influence of other variables such as age or sex on the outcome.

Inclusion/exclusion criteria. Pre-established criteria for excluding mice from behavioural analysis included (1) missed injections, (2) anomalies during behavioural testing, such as mice falling off the elevated plus maze, (3) damage to or loss of optical fibres, (4) statistical outliers, as determined by the Grubb's test.

Sample size. A power analysis was used to determine the ideal sample size for behaviour experiments. Assuming a normal distribution, a 20% change in mean and 15% variation, we determined that we would need 8 mice per group. In some cases, mice were excluded due to missed injections or lost optical fibres resulting in fewer than 8 mice per group. For electrophysiology experiments, we aimed for 5–7 cells from 3–4 mice.

Statistics. Data are presented as means ± s.e.m. For comparisons with only two groups, P values were calculated using paired or unpaired t -tests as described in the figure legends. Comparisons across more than two groups were made using a one-way ANOVA, and a two-way ANOVA was used when there was more than one independent variable. A Bonferroni post-test was used following significance with an ANOVA. In cases in which ANOVA was used, the data met the assumptions of equality of variance and independence of cases. If the condition of equal variances was not met, Welch's correction was used. Some of the sample groups were too small to detect normality (<8 samples) but parametric tests were used because nonparametric tests lack sufficient power to detect differences in small samples (Graphpad Statistics Guide – <http://www.graphpad.com>). The standard error of the mean is indicated by error bars for each group of data. Differences were considered significant at P values below 0.05. All data were analysed with GraphPad Prism software.

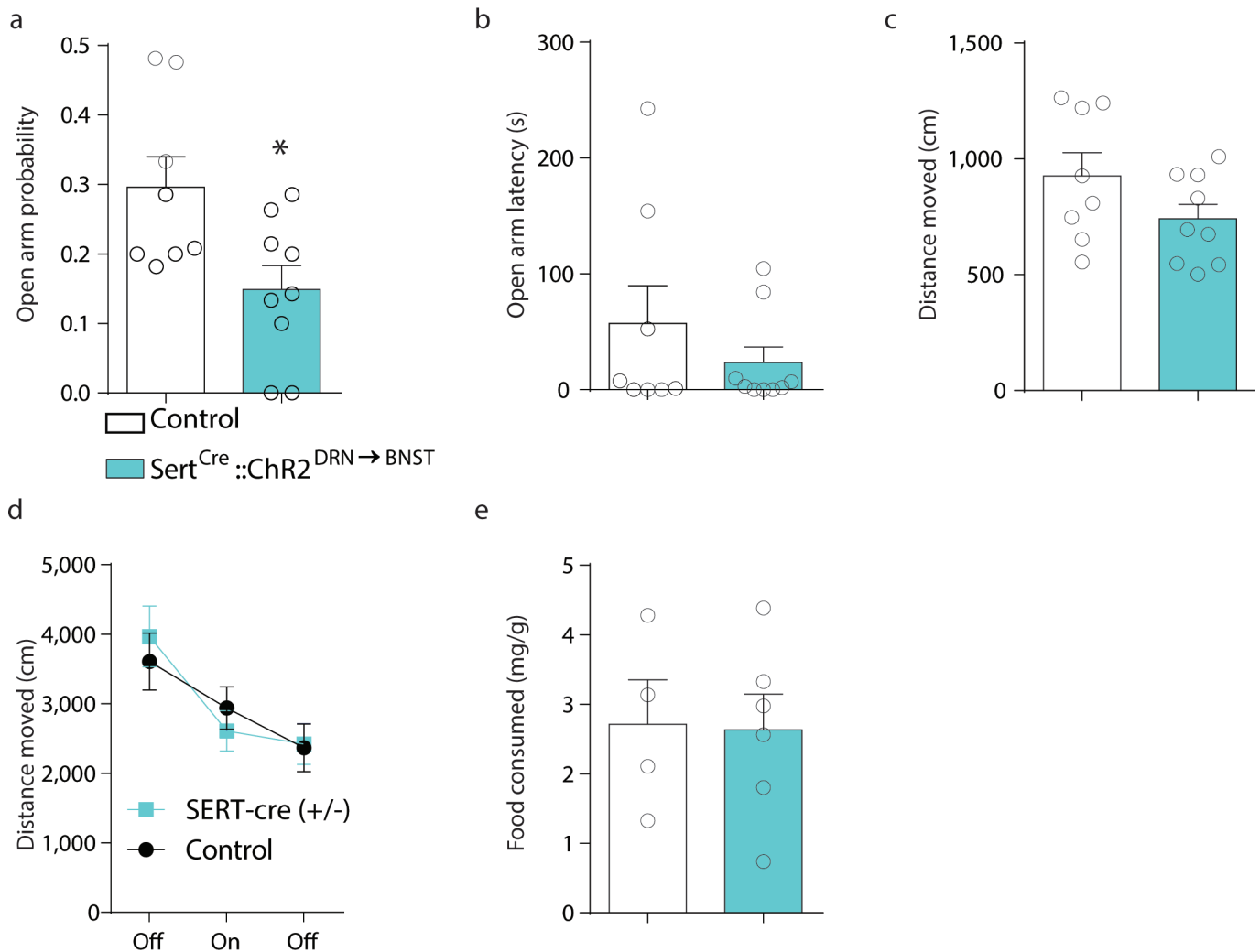
21. Krashes, M. J. *et al.* An excitatory paraventricular nucleus to AgRP neuron circuit that drives hunger. *Nature* **507**, 238–242 (2014).
22. Madisen, L. *et al.* A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nature Neurosci.* **13**, 133–140 (2010).
23. Bath, B. D. *et al.* Subsecond adsorption and desorption of dopamine at carbon-fiber microelectrodes. *Anal. Chem.* **72**, 5994–6002 (2000).
24. Hefner, K. *et al.* Impaired fear extinction learning and cortico-amygdala circuit abnormalities in a common genetic mouse strain. *J. Neurosci.* **28**, 8074–8085 (2008).



Extended Data Figure 1 | See next page for caption.

Extended Data Figure 1 | *In vivo* recordings in BNST neurons during fear conditioning reveal opposite patterns of activation during acquisition and recall. **a, b**, Representative neuronal firing rate (**a**) and population Z score of the firing rate (**b**) for BNST neurons ($n = 45$ cells from 7 mice) 30 s before conditioned stimulus (tone), during the conditioned stimulus (CS), and 30 s after the unconditioned stimulus. **c**, Percentage time spent freezing during fear acquisition, cued fear recall and contextual fear recall. **d**, Electrode placements for BNST recordings. **e**, Raw firing rates during freezing (blue) versus movement (red) epochs were averaged across all putative principal neurons (firing rate < 10 Hz). Acquisition: cells in BNST exhibited greater average firing rates during freezing epochs compared to movement epochs during CS3 ($t_{44} = 2.88$, $P < 0.01$, Student's unpaired two-tailed t -test), CS4 ($t_{44} = 3.14$, $P < 0.01$, Student's unpaired two-tailed t -test), and CS5 ($t_{44} = 4.4$, $P < 0.001$, Student's unpaired two-tailed t -test) ($n = 45$ cells from 7 mice). CS recall: average firing rates during freezing epochs decreased over CS presentations such that firing during block 5 was significantly less than block 1 ($t_{41} = 3.44$, $P = 0.001$, Student's unpaired two-tailed t -test). Freezing firing rates during block 5 were also significantly less than movement epochs during block 5 ($t_{41} = 4.03$, $P < 0.001$, Student's unpaired two-tailed t -test) ($n = 42$ cells from 7 mice). CX test: average

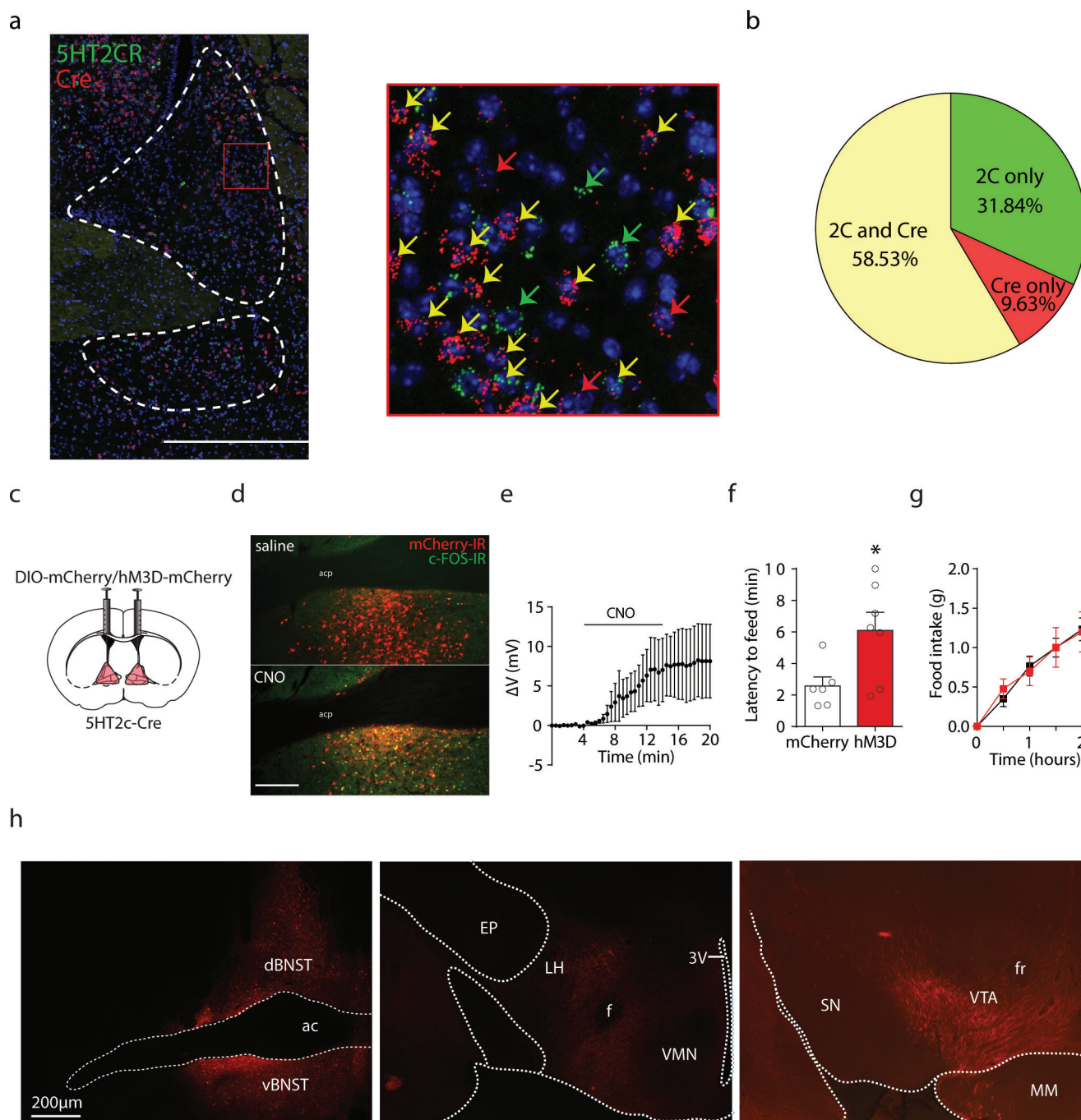
firing rate was significantly greater during movement versus freezing epochs during minute 1 ($t_{44} = 4.83$, $P < 0.001$, Student's unpaired two-tailed t -test), minute 2 ($t_{44} = 3.17$, $P < 0.01$, Student's unpaired two-tailed t -test), and minute 5 ($t_{44} = 4.36$, $P < 0.001$, Student's unpaired two-tailed t -test) ($n = 45$ cells from 7 mice). **f**, Freezing-related changes in firing rates during the CS were determined by measuring the ratio of average firing rates during freezing versus movement epochs for each session. Acquisition: activity during freezing epochs increased significantly relative to movement epochs during CS4 ($t_{45} = 3.26$, $P < 0.01$, Student's unpaired two-tailed t -test) and CS5 ($t_{45} = 2.17$, $P < 0.05$, Student's unpaired two-tailed t -test) ($n = 46$ cells from 7 mice). CS recall: freezing significantly suppressed activity relative to movement epochs during the last two CS presentations ($t_{47} = 5.29$, $P < 0.001$, Student's unpaired two-tailed t -test) ($n = 48$ cells from 7 mice). CX test: freezing significantly suppressed activity during minutes 1 ($t_{44} = 6.06$, $P < 0.001$, Student's unpaired two-tailed t -test), minute 2 ($t_{44} = 2.92$, $P < 0.01$, Student's unpaired two-tailed t -test), and minute 5 ($t_{44} = 3.55$, $P = .001$, Student's unpaired two-tailed t -test) ($n = 45$ cells from 7 mice). **g**, Plots showing correlation between freezing behaviour and firing rate of BNST neurons across sessions and for all sessions. Data are mean \pm s.e.m. * $P < 0.05$ ** $P < 0.01$; *** $P < 0.001$. Scale bar, 100 μ m.



Extended Data Figure 2 | Effects of optogenetic stimulation of 5-HT inputs to the BNST on feeding, anxiety and locomotion.

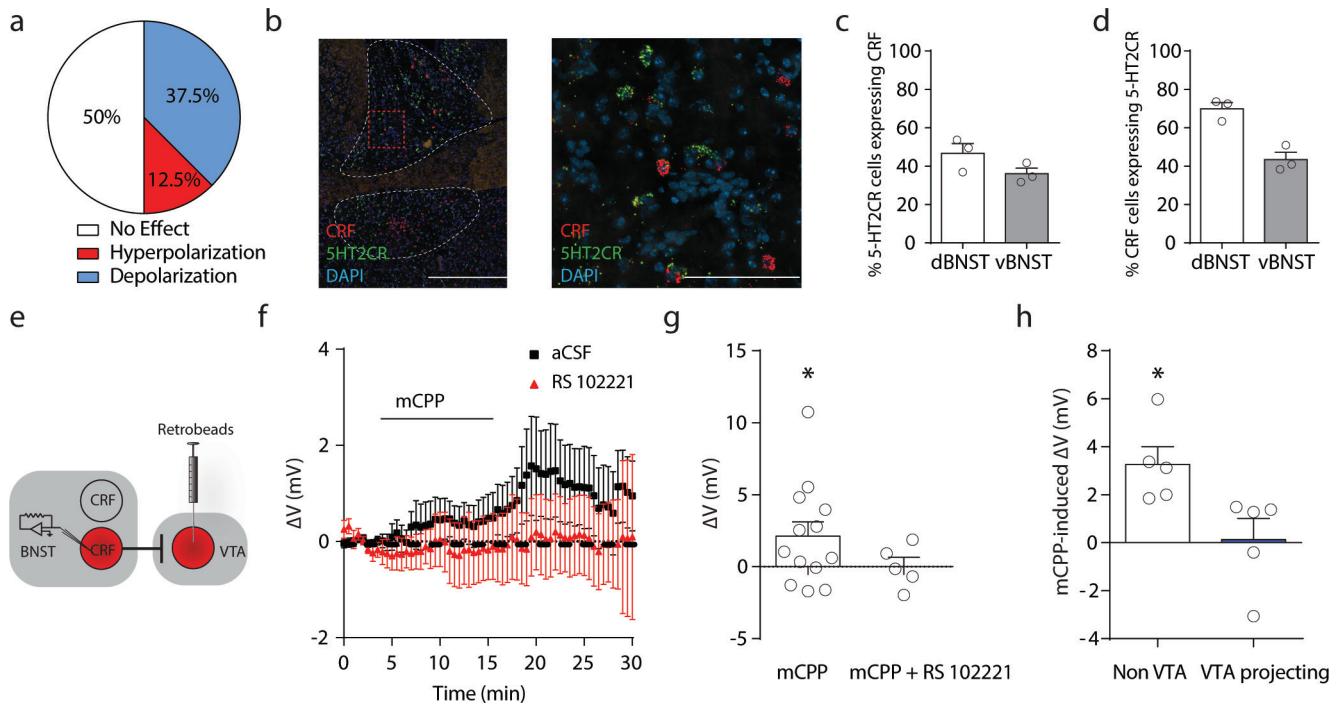
a–c, Sert^{Cre}::ChR2^{DRN→BNST} mice exhibited reduced probability ($t_{15} = 2.67$, $P < 0.05$, Student's unpaired two-tailed t -test, $n = 8$ control, $n = 9$ ChR2) and latency ($t_{15} = 1.003$, $P > 0.05$, Student's unpaired two-tailed t -test,

$n = 8$ control, $n = 9$ ChR2) to enter the open arms of the EPM without exhibiting locomotor deficits. **d**, **e**, Photostimulation of 5-HT^{DRN→BNST} terminals had no effect on locomotor activity in the open field (**d**) ($n = 9$ control, $n = 11$ ChR2) or home cage feeding (**e**) ($n = 4$ control, $n = 6$ ChR2). Data are mean \pm s.e.m. * $P < 0.05$.



Extended Data Figure 3 | Chemogenetic activation of 5-HT_{2C}R-expressing neurons in the BNST increases anxiety-like behaviour. **a**, Confocal images of coronal BNST slices obtained from *Htr2c^{Cre}* mice following double fluorescence *in situ* hybridization for 5-HT_{2C}R and Cre. Yellow arrows indicate cells in which there is co-localization, red arrows indicate cells in which only Cre is expressed and green arrows indicate cells in which only 5-HT_{2C}R is expressed. **b**, Pie chart representing the distribution of genetic markers in BNST neurons. **c**, Experimental configuration in *Htr2c^{Cre}::hM3Dq^{BNST}* mice. **d**, Coronal images showing c-fos induction in 5-HT_{2C}R expressing neurons in the BNST of *Htr2c^{Cre}::hM3Dq^{BNST}* or *Htr2c^{Cre}::mCherry^{BNST}* mice

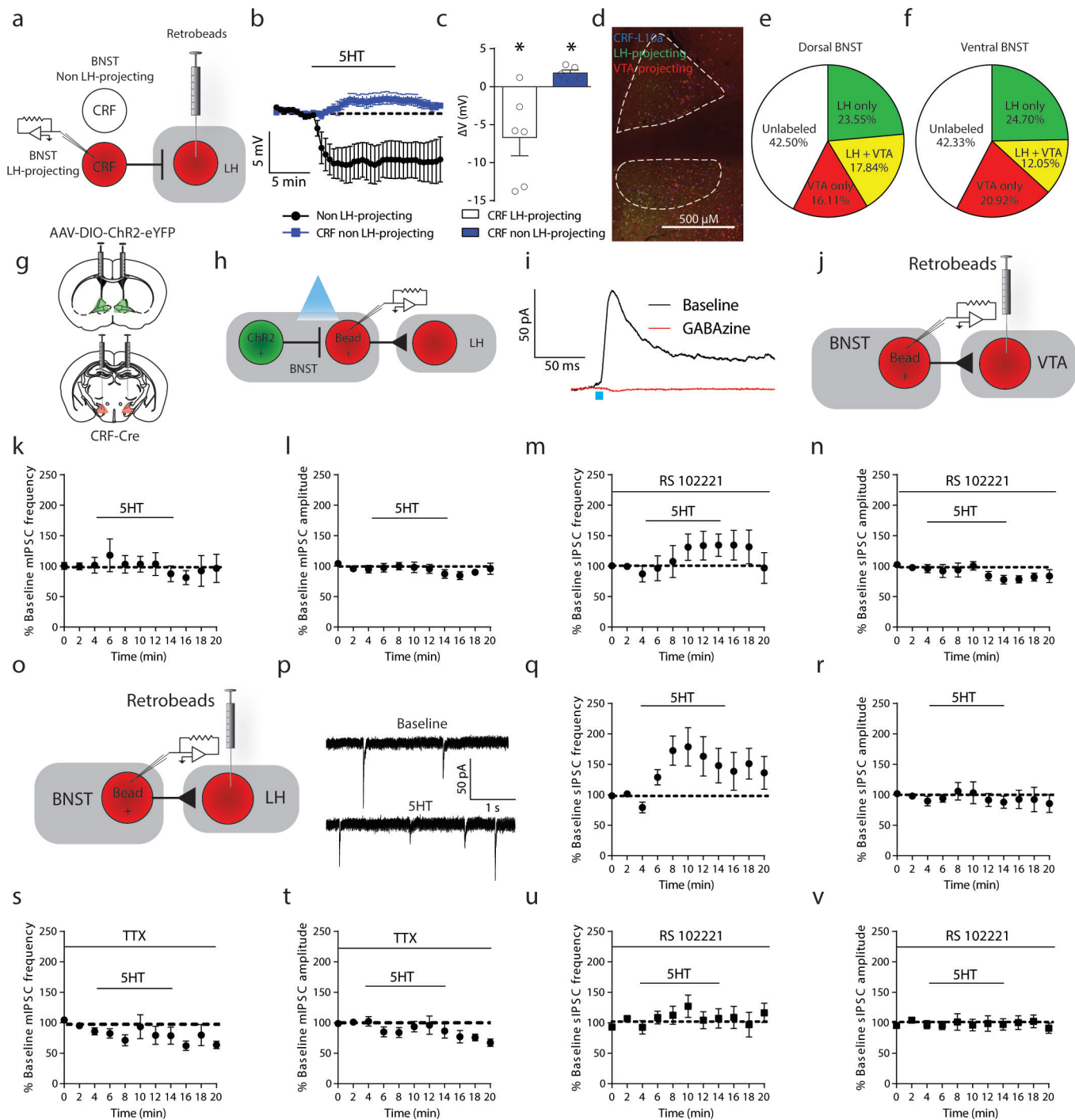
following CNO injection. **e**, Bath application of CNO depolarized 5-HT_{2C}R-expressing neurons expressing hM3Dq in slice ($n = 3$ cells from 3 mice). **f**, Chemogenetic stimulation of 5-HT_{2C}R expressing neurons in BNST increased latency to feed in the NSF ($t_{11} = 2.591$, $P < 0.05$, Student's unpaired two-tailed t -test, $n = 6$; mCherry, $n = 7$ hM3Dq). **g**, Chemogenetic activation of 5-HT_{2C}R-expressing BNST neurons had no effect on home cage feeding ($n = 5$ mCherry, $n = 6$ hM3Dq). **h**, Confocal images from *Htr2c^{Cre}::mCherry^{BNST}* mice showing mCherry expression in 5-HT_{2C}R-expressing soma in the BNST and fibres in the LH and VTA. Data are mean \pm s.e.m. * $P < 0.05$. Scale bar, 100 μ m.



Extended Data Figure 4 | Electrophysiological characterization of 5-HT responses and 5-HT receptor expression in CRF^{BNST} neurons.

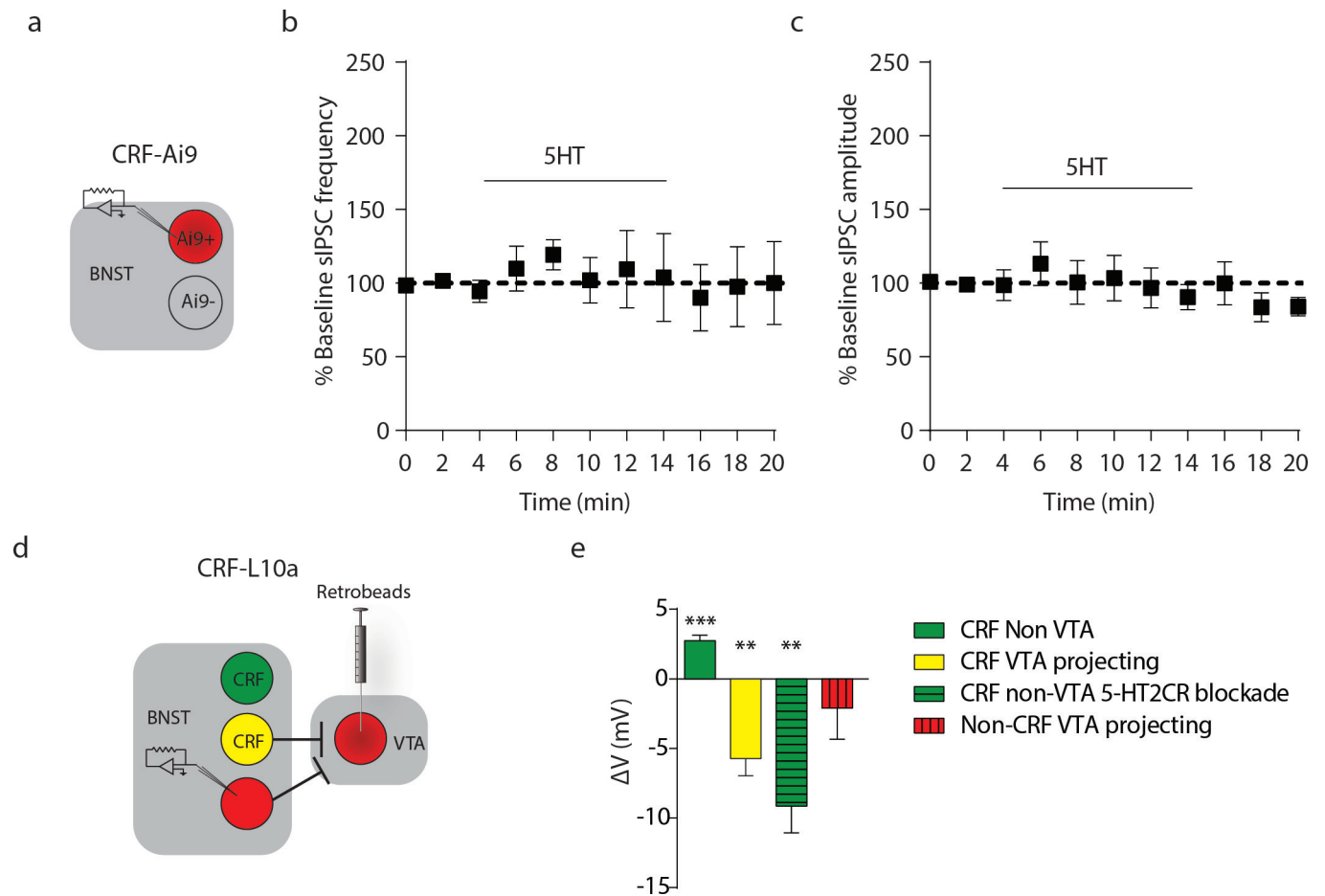
a, A pie chart showing the distribution of CRF^{BNST} neurons that were depolarized, hyperpolarized, or had no response to 5-HT ($n=8$ cells from 4 mice). **b**, Coronal images of the BNST showing co-localization of 5-HT_{2C}Rs with CRF mRNA using double fluorescence *in situ* hybridization. **c**, **d**, Histograms showing the percentage of 5-HT_{2C} neurons that express CRF and the percentage of CRF neurons that express 5-HT_{2C}Rs in the BNST ($n=3$ slices from 3 mice). **e**, Recording configuration in CRF^{BNST} neurons. **f**, Slice electrophysiology in BNST of *Crf* reporter mice showing

depolarization of all (VTA-projecting and non-projecting) CRF neurons following bath application of the 5-HT₂ receptor agonist mCPP ($n=12$ cells from 6 mice) and blockade of this response by the 5-HT_{2C} receptor antagonist RS-102221 ($n=5$ cells from 3 mice). **g**, Change in membrane potential induced by mCPP ($t_{12}=2.18$, $P<0.05$, one-sample *t*-test, $n=13$ cells from 6 mice) is blocked by a 5-HT_{2C}R antagonist ($n=5$ cells from 3 mice). **h**, mCPP selectively depolarizes non-VTA-projecting CRF^{BNST} neurons ($n=5$ cells from 2 mice non-VTA-projecting CRF, $n=5$ cells from 4 mice VTA-projecting CRF). Data are mean \pm s.e.m. * $P<0.05$.



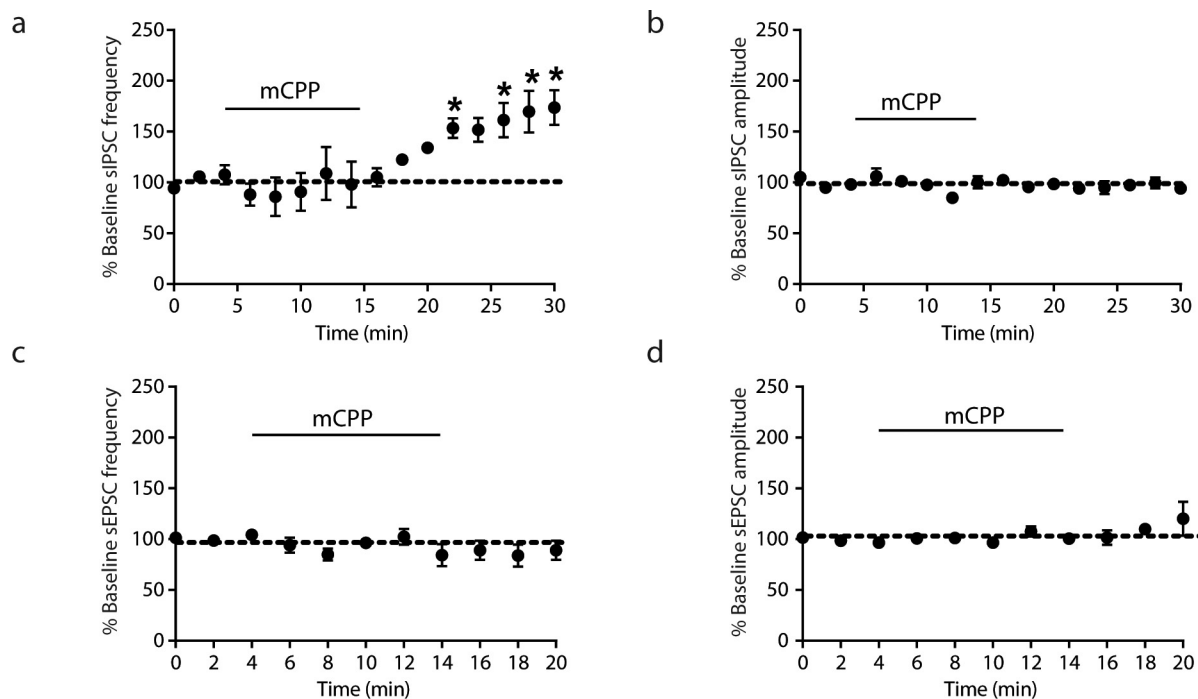
Extended Data Figure 5 | 5-HT activates inhibitory microcircuits in the BNST that modulate outputs to the LH. **a**, Recording configuration in CRF reporter mice infused with retrograde tracer beads in the LH. **b**, Average traces of 5-HT induced depolarization in LH projecting versus non-projecting neurons. **c**, Histograms showing 5-HT induced depolarization in non-LH projecting BNST neurons ($t_4 = 4.425$, $P < 0.05$, one-sample t -test, $n = 5$ cells from 3 mice) and hyperpolarization in LH-projecting neurons ($t_5 = 2.789$, $P < 0.05$, one-sample t -test, $n = 6$ cells from 3 mice). **d**, Confocal image of retrogradely CTB-labelled VTA (red) and LH (green) outputs in a *CRF-L10a* reporter (blue). **e, f**, Pie charts depicting the percentage of LH-projecting only, VTA-projecting only, collateralizing, and CTB-negative (unlabelled) CRF in neurons in the dorsal and ventral aspects of the BNST ($n = 6$ hemispheres from 3 mice). **g**, Experimental schematic depicting viral infusions into the BNST and retrograde tracer bead infusions into the LH of *Crf^{Cre}::ChR2^{BNST}* mice. **h**, Recording configuration in *Crf^{Cre}::ChR2^{BNST}* mice with LH tracer beads.

i, Representative trace of light evoked IPSCs in LH-projecting neurons ($n = 7$ cells from 4 mice) and blockade of this light evoked response by GABAzine ($n = 2$ cells from 2 mice). **j**, Recording configuration in VTA-projecting neurons in the BNST of C57BL/6 mice. **k, l**, 5-HT has no effect on miniature IPSC frequency or amplitude in BNST→VTA projecting neurons ($n = 7$ from 4 mice). **m, n**, 5-HT has no effect on sIPSC frequency or amplitude in the presence of the 5-HT_{2C}R antagonist RS-102221 ($n = 5$ cells from 4 mice). **o**, Recording configuration in LH projecting neurons in the BNST of C57BL/6 mice. **p**, Representative traces showing an increase in sIPSC frequency in the presence of 5-HT for 6 cells from 3 mice. **q, r**, 5-HT increases sIPSC frequency but not amplitude in BNST→LH projecting neurons ($F_{11,55} = 11.65$, $P < 0.01$, repeated measures one-way ANOVA, $n = 6$ cells from 3 mice). **s, t**, 5-HT has no effect on miniature IPSC frequency or amplitude ($n = 5$ cells from 3 mice). **u, v**, 5-HT has no effect on sIPSC frequency or amplitude in the presence of RS-102221 ($n = 6$ cells from 4 mice). Data are mean \pm s.e.m. * $P < 0.05$.



Extended Data Figure 6 | 5-HT does not alter GABAergic transmission in CRF neurons nor does it directly excite non-CRF VTA-projecting neurons in the BNST. **a**, Recording configuration in CRF^{BNST} neurons in a CRF reporter. **b**, **c**, 5-HT has no effect on sIPSC frequency or amplitude in the total population of CRF neurons ($n = 5$ cells from 3 mice). **d**, Recording configuration in non-CRF, VTA-projecting neurons in the BNST and average trace of 5-HT effect on membrane potential

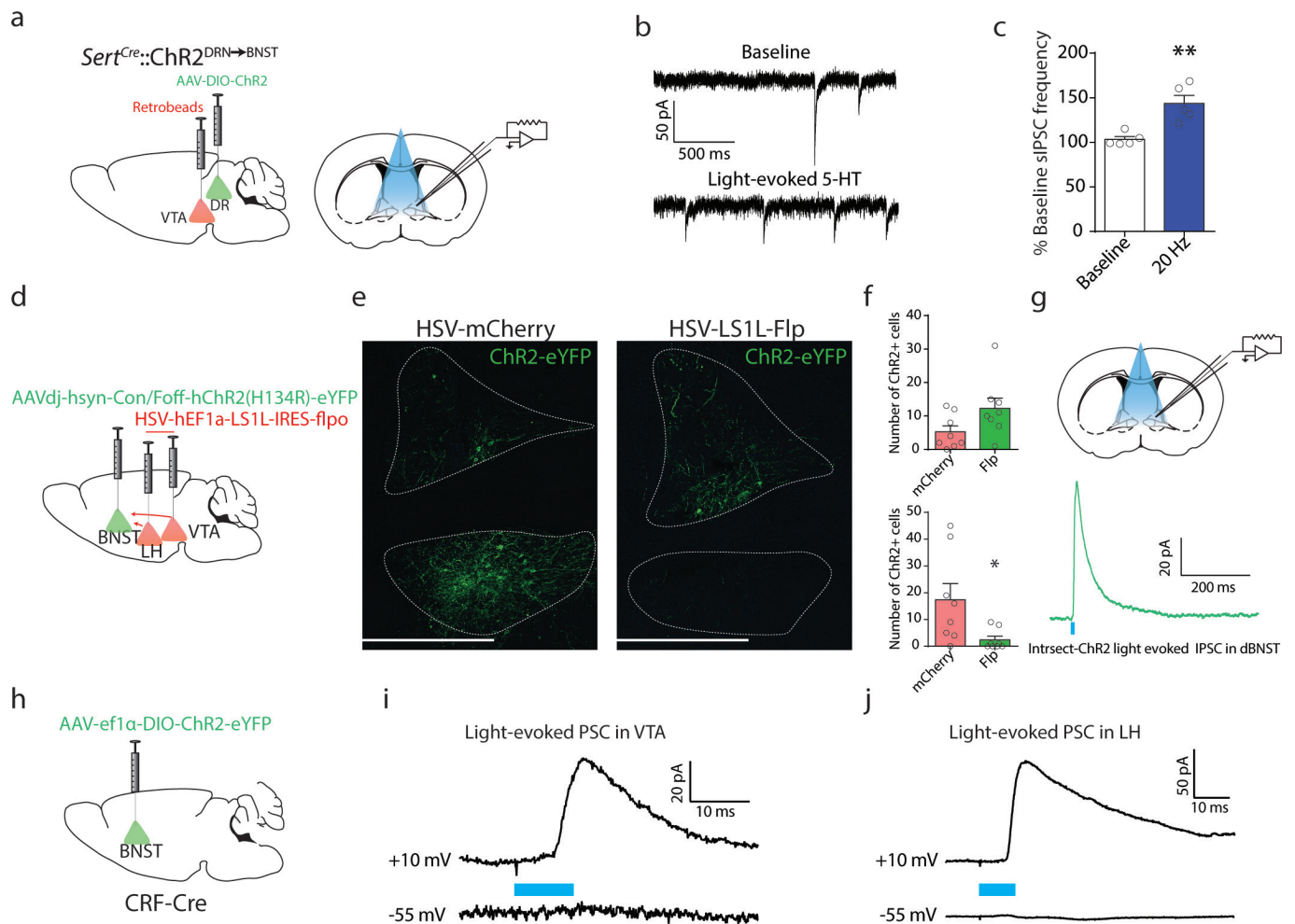
in non-CRF, VTA-projecting neurons in the presence of tetrodotoxin. **e**, Histogram summarizing 5-HT effects on membrane potential in local and VTA-projecting CRF neurons and local CRF neurons in the presence of the 5-HT_{2C} receptor antagonist RS-102221 (same data shown in Fig. 2b) juxtaposed with the lack of effect of 5-HT on membrane potential in non-CRF, VTA-projecting neurons ($t_4 = 0.9381$, ns, one-sample t -test, $n = 5$ cells from 3 mice). Data are mean \pm s.e.m. ** $P < 0.01$; *** $P < 0.001$.



Extended Data Figure 7 | The 5-HT₂ agonist mCPP increases GABAergic but not glutamatergic transmission in the BNST.

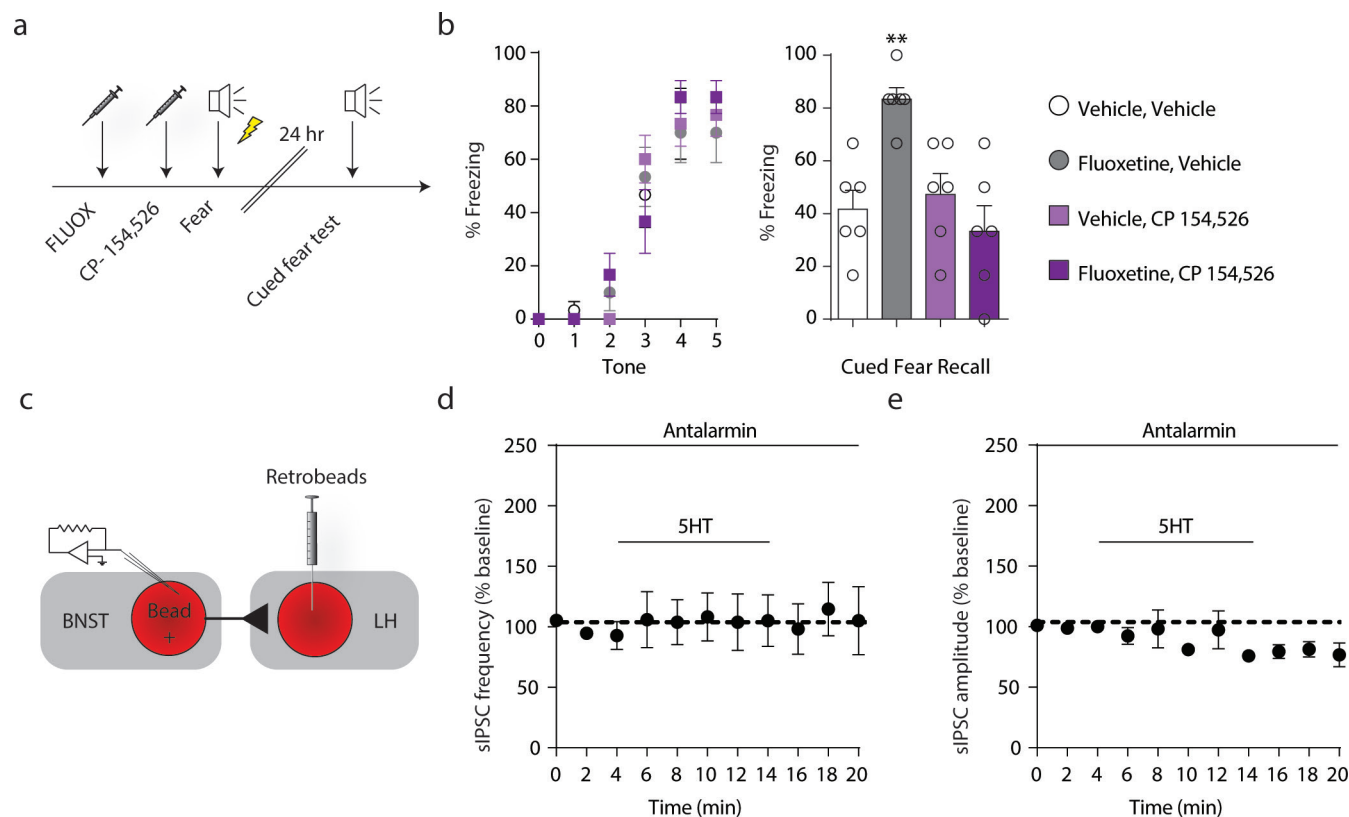
a, b, mCPP increases sIPSC frequency ($F_{15,30} = 1.863$, $P < 0.001$, Repeated measures one-way ANOVA, $n = 3$ cells from 3 mice) but not amplitude

in the BNST of C57BL/6 mice. **c, d,** mCPP has no effect on spontaneous excitatory postsynaptic current (sEPSC) frequency or amplitude in the BNST of C57BL/6 mice ($n = 5$ cells from 3 mice). Data are mean \pm s.e.m. * $P < 0.05$.



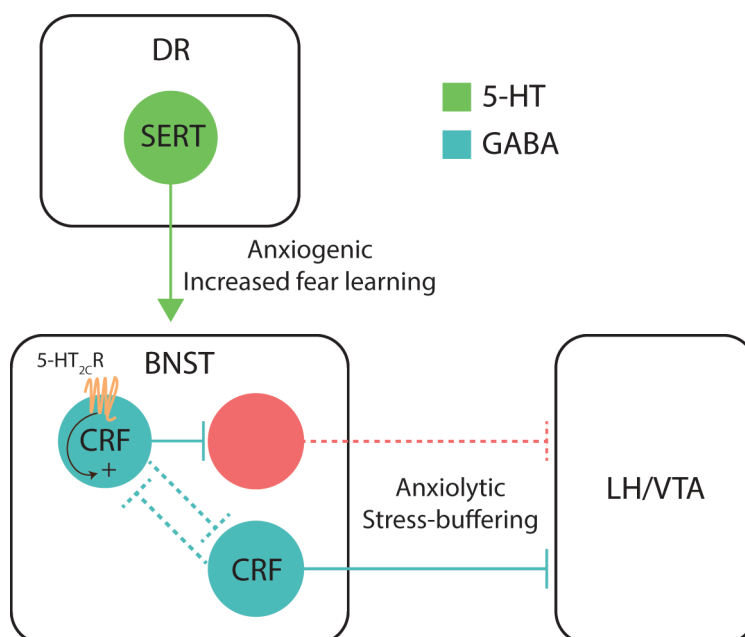
Extended Data Figure 8 | Optogenetic and intrasectional characterization of 5-HT-CRF circuits in the BNST and outputs to the midbrain. **a**, Experimental design and recording configuration from *Sert^{Cre}::Chr2^{DRN→BNST}* mouse with retrograde tracer beads in the VTA. **b**, Representative traces for 5 cells from 3 mice depicting the increase in sIPSCs in VTA-projecting neurons in the BNST following light-evoked 5-HT release on sIPSC frequency in VTA-projecting neurons ($t_4 = 4.890$, $P < 0.01$, one-sample t -test, $n = 5$ cells from 3 mice). **d**, Experimental configuration in *Cr^fCre*::Intrsect-ChR2^{BNST} mice. **e**, Representative images from 4 *Cr^fCre*::HSV-LSL1-mCherry-flpo^{VTA/LH} mice and 4 *Cr^fCre*::HSV-LSL1-mCherry^{VTA/LH} mice injected with Intrsect-ChR2-eYFP in the

BNST. **f**, Cell counts of eYFP⁺ neurons from HSV-LSL1-flpo and HSV-LSL1-mCherry injected *Cr^fCre*::Intrsect-ChR2^{BNST} mice indicating the number of non-projecting CRF neurons compared to the total CRF population in the dorsal (top panel; $t_{14} = 1.959$, ns, Student's unpaired two-tailed t -test, $n = 4$ mice, 8 hemispheres per group) and ventral aspects of the BNST (bottom panel; $t_7 = 2.431$, $P < 0.05$, Student's unpaired Welch's corrected two-tailed t -test, $n = 4$ mice, 8 hemispheres per group). **g**, Recording configuration and light-evoked IPSC showing local GABA release from non-projecting CRF neurons in the BNST. **h**, Stereotaxic injection of ChR2 in *Cr^fCre* mouse. **i**, **j**, Light evoked IPSCs in the VTA and LH indicating that CRF projections to these regions are GABAergic. Data are mean \pm s.e.m. * $P < 0.05$; ** $P < 0.01$.



Extended Data Figure 9 | Pharmacological blockade of CRF_1 receptors reduces fluoxetine-induced aversive behaviour and 5-HT enhancement of GABAergic transmission in the BNST. **a**, Experimental schedule of injections and behaviour. **b**, CRF_1 R antagonist does not modify fear acquisition but reduces fluoxetine enhancement of cued fear recall ($F_{1,20} = 13.70$, $P < 0.01$, two-way ANOVA, $n = 6$ per group). **c**, Recording configuration in BNST neurons that project to the LH in C57BL/6 mice.

d, Bath application of a CRF_1 R antagonist blocks the 5-HT induced increase in sIPSC frequency in LH-projecting neurons in the BNST ($F_{10,30} = 0.2213$, ns, Repeated measures one-way ANOVA, $n = 4$ cells from 2 mice). **e**, There was a reduction in sIPSC amplitude during 5-HT bath application and CRF_1 R blockade ($F_{10,30} = 2.941$, $P < 0.05$, Repeated measures one-way ANOVA, $n = 4$ cells from 2 mice). Data are mean \pm s.e.m. ** $P < 0.01$.



Extended Data Figure 10 | Model of a serotonin-sensitive inhibitory microcircuit in the BNST that modulates anxiety and aversive learning. Serotonin inputs to the BNST activate 5-HT_{2C}Rs expressed in non-projecting 'local' CRF neurons. These local CRF neurons promote anxiety and fear by inhibiting anxiolytic outputs to the VTA and LH that are putatively GABAergic. Another discrete subset of CRF neurons, which are

inhibited by 5-HT, send direct, inhibitory projections to the VTA and LH. These CRF^{BNST} output neurons are GABAergic and putatively anxiolytic and stress buffering. Blue dashed lines indicate hypothesized additional synapses between CRF^{BNST} neurons. Dashed red line indicates a putatively GABAergic synapse.

HER2 expression identifies dynamic functional states within circulating breast cancer cells

Nicole Vincent Jordan¹, Aditya Bardia^{1,2}, Ben S. Wittner^{1,2}, Cyril Benes^{1,2}, Matteo Ligorio^{1,3}, Yu Zheng¹, Min Yu^{1†}, Tilak K. Sundaresan^{1,2}, Joseph A. Licausi¹, Rushil Desai¹, Ryan M. O'Keefe¹, Richard Y. Ebright¹, Myriam Boukhali¹, Srinjoy Sil¹, Maristela L. Onozato^{1,4}, Anthony J. Iafrate^{1,4}, Ravi Kapur⁵, Dennis Sgroi^{1,4}, David T. Ting^{1,2}, Mehmet Toner^{3,5}, Sridhar Ramaswamy^{1,2}, Wilhelm Haas^{1,2}, Shyamala Maheswaran^{1,3} & Daniel A. Haber^{1,2,6}

Circulating tumour cells in women with advanced oestrogen-receptor (ER)-positive/human epidermal growth factor receptor 2 (HER2)-negative breast cancer acquire a HER2-positive subpopulation after multiple courses of therapy^{1,2}. In contrast to HER2-amplified primary breast cancer, which is highly sensitive to HER2-targeted therapy, the clinical significance of acquired HER2 heterogeneity during the evolution of metastatic breast cancer is unknown. Here we analyse circulating tumour cells from 19 women with ER⁺/HER2⁻ primary tumours, 84% of whom had acquired circulating tumour cells expressing HER2. Cultured circulating tumour cells maintain discrete HER2⁺ and HER2⁻ subpopulations: HER2⁺ circulating tumour cells are more proliferative but not addicted to HER2, consistent with activation of multiple signalling pathways; HER2⁻ circulating tumour cells show activation of Notch and DNA damage pathways, exhibiting resistance to cytotoxic chemotherapy, but sensitivity to Notch inhibition. HER2⁺ and HER2⁻ circulating tumour cells interconvert spontaneously, with cells of one phenotype producing daughters of the opposite within four cell doublings. Although HER2⁺ and HER2⁻ circulating tumour cells have comparable tumour initiating potential, differential proliferation favours the HER2⁺ state, while oxidative stress or cytotoxic chemotherapy enhances transition to the HER2⁻ phenotype. Simultaneous treatment with paclitaxel and Notch inhibitors achieves sustained suppression of tumorigenesis in orthotopic circulating tumour cell-derived tumour models. Together, these results point to distinct yet interconverting phenotypes within patient-derived circulating tumour cells, contributing to progression of breast cancer and acquisition of drug resistance.

We documented the emergence of HER2⁺ circulating tumour cells (CTCs) in patients initially diagnosed with ER-positive/HER2-negative (ER⁺/HER2⁻) breast cancer, after multiple courses of therapy for recurrent metastatic breast cancer. Using microfluidic CTC-iChip purification followed by imaging flow cytometry³, 16 out of 19 (84%) patients had HER2⁺ CTCs (Fig. 1a, Extended Data Fig. 1a and Supplementary Table 1). Twenty-two individual CTCs from two representative patients (Brx-42, Brx-82) were isolated and subjected to single-cell RNA sequencing (scRNA-seq). HER2 expression was bimodal in distribution (≤ 1 read per million (RPM) versus median 133, range 32–217 RPM; $P = 7.5 \times 10^{-6}$) (Fig. 1b), indicating the existence of discrete HER2⁺ and HER2⁻ subpopulations. In these patients, the fraction of HER2⁺ CTCs increased with disease progression (Extended Data Fig. 1b). HER2⁺ CTCs were not restricted to ER⁺/HER2⁻ breast cancer: 2 out of 13 patients with ER⁻/PR⁻/HER2⁻ (triple negative) breast cancer also had HER2⁺ and HER2⁻ CTC subpopulations (Extended Data Fig. 1c). In ER⁺/HER2⁻ breast cancers, immunohistochemical (IHC) staining

of patient-matched metastatic tumour biopsies showed increased HER2⁺ staining, compared with primary tumours (Fig. 1c). Unlike HER2-amplified breast cancer, HER2⁺ tumour cells within metastatic lesions did not have evidence of gene amplification (Extended Data Fig. 1d).

The CTC-iChip efficiently captures viable CTCs³, enabling derivation of CTC cultures⁴. We established CTC lines (Brx-42, Brx-82, Brx-142) with discrete HER2⁺/HER2⁻ subpopulations comparable to patient-matched primary CTCs (Fig. 1a, d and Extended Data Fig. 1e, f). Acquired HER2 expression was not due to gene amplification, and no distinguishing mutations were identified between HER2⁺ and HER2⁻ subpopulations (Extended Data Fig. 1g and Supplementary Table 2). Fluorescence-activated cell sorting (FACS) of HER2⁺ versus HER2⁻ subpopulations showed distinct functional properties: HER2⁺ CTCs had a higher proliferation rate (Fig. 1e), with increased staining for the proliferation marker Ki67, but no change in apoptotic markers cleaved-caspase 3 or annexin 5 (Extended Data Fig. 2a, b).

We tested the relative tumorigenicity of HER2⁺ versus HER2⁻ CTCs following injection into the mouse mammary fat pad. Both FACS-purified HER2⁺ and HER2⁻ CTCs generated tumours, with HER2⁺ tumours being larger and having a higher frequency of lung metastases (Fig. 1f and Extended Data Fig. 2c, d). Despite differences in proliferation, limiting dilution studies showed that HER2⁺ and HER2⁻ CTCs initiate tumours from as few as 200 cells, pointing to comparable progenitor potential (Extended Data Fig. 2e).

The coexistence of HER2⁺ and HER2⁻ CTCs, despite differing proliferation rates, led us to test whether these subpopulations are capable of interconversion. After 4 weeks in culture, FACS-purified green fluorescent protein (GFP)-tagged HER2⁻ CTCs acquired HER2⁺ cells (Brx-82: 42%; Brx-142: 46%), while HER2⁺ CTCs generated HER2⁻ cells at lower efficiency (Brx-82: 5%; Brx-142: 11%) (Fig. 2a, b and Extended Data Fig. 3a). By 8 weeks, the parental HER2⁺/HER2⁻ composition was nearly re-established (Fig. 2b). This interconversion was also evident by mixing equal proportions of GFP⁺/HER2⁺ and GFP⁻/HER2⁻ CTCs, with the emergence of GFP⁺/HER2⁻ and GFP⁻/HER2⁺ cells, respectively (Extended Data Fig. 3a).

To better define the timing of HER2⁺/HER2⁻ interconversion, we established single-cell-derived CTC colonies using HER2-based FACS, followed by sequential confocal microscopy. Colonies were scored for HER2 and EpCAM expression at 1-, 3-, 5- to 9-, 10- to 19- and >20-cell stages. Single HER2⁻ CTCs initially proliferated slowly (Extended Data Fig. 3b), and first acquired HER2⁺ daughter cells at the 5- to 9-cell stage (6.5%), with rapid interconversion thereafter (10–19 cells: 47%, >20 cells: 59%; Fig. 2c, d). The more rapidly proliferating single HER2⁺ CTCs also generated HER2⁻ progeny at the 5- to 9-cell stage (5%), but the proportion of HER2⁻ CTCs rose more slowly

¹Massachusetts General Hospital Cancer Center, Harvard Medical School, Charlestown, Massachusetts 02129, USA. ²Department of Medicine, Harvard Medical School, Boston, Massachusetts 02114, USA. ³Department of Surgery, Harvard Medical School, Boston, Massachusetts 02114, USA. ⁴Department of Pathology, Harvard Medical School, Boston, Massachusetts 02114, USA. ⁵Center for Bioengineering in Medicine and Shriners Hospital, Harvard Medical School, Boston, Massachusetts 02114, USA. ⁶Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA. †Present address: Department of Stem Cell Biology and Regenerative Medicine, University of Southern California, California 90033, USA.

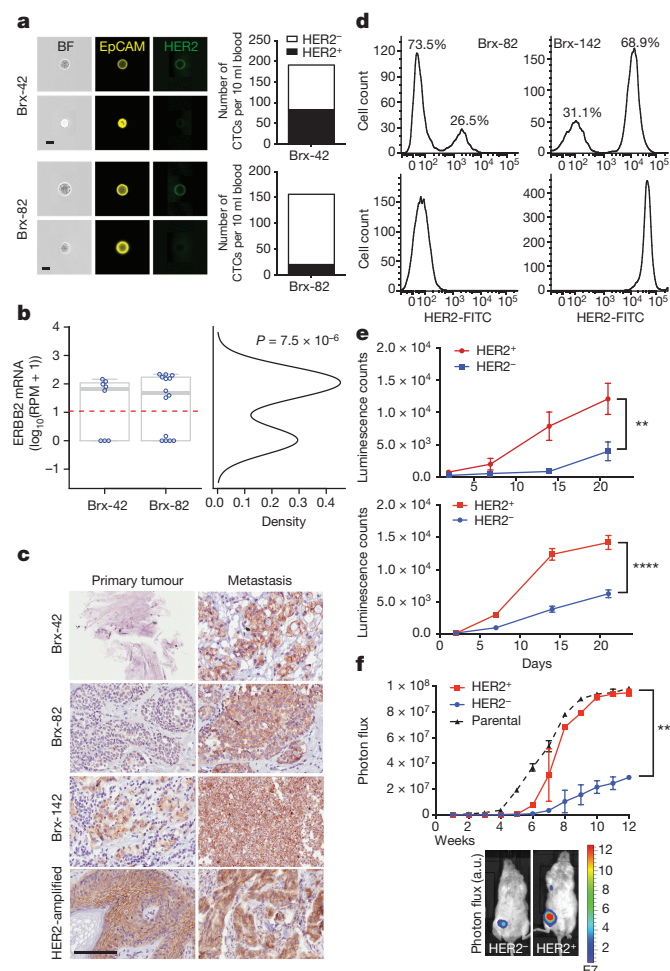


Figure 1 | Distinct properties of $HER2^{+}$ and $HER2^{-}$ CTC subpopulations from patients with advanced $ER^{+}/HER2^{-}$ breast cancer. **a**, Quantitation by imaging flow cytometry of $HER2^{+}$ and $HER2^{-}$ CTCs isolated from patients Brx-42, Brx-82. EpCAM (yellow) and HER2 (green). Scale bar, 10 μ m. **b**, Bimodal distribution of *ERBB2* RNA-seq reads from single CTCs, Hartigan's dip test, $P = 7.5 \times 10^{-6}$, $n = 22$ ($HER2^{-} \leq 1$ RPM; $HER2^{+} > 133$, range 32–217). **c**, IHC for HER2 (brown) in matched metastatic versus primary tumours (Brx-42, Brx-82, Brx-142) compared with HER2-amplified tumour (control). Scale bar, 100 μ m; tumour data (Supplementary Table 1). **d**, FACS of cultured CTCs, showing discrete $HER2^{+}$ and $HER2^{-}$ subpopulations. MDA-231 (triple-negative breast cancer (TNBC)) and SKBR3 (HER2-amplified) cells are shown as control. **e**, Differential proliferation of FACS-purified $HER2^{+}$ (red) and $HER2^{-}$ (blue) subpopulations from cultured CTCs; two-way analysis of variance (ANOVA) $P < 0.01$ (Brx-82), $P < 0.0001$ (Brx-142); $n = 6$; s.d. (error bar). **f**, Increased *in vivo* growth of orthotopic mammary tumours derived from FACS-purified, $HER2^{+}$ CTCs compared with $HER2^{-}$ cells; $n = 8$; two-way ANOVA, $P < 0.0001$; s.d. (error bar).

(10–19 cells: 17%, >20 cells: 22%; Fig. 2c, d). Thus, interconversion between $HER2^{+}/HER2^{-}$ phenotypes occurs spontaneously as early as four cell doublings.

Interconversion between $HER2^{+}$ and $HER2^{-}$ phenotypes was also tested *in vivo* by orthotopic inoculation of FACS-purified cultured CTCs. Tumours established from $HER2^{-}$ CTCs displayed $HER2^{+}$ subpopulations, and vice versa (Fig. 2e and Extended Data Fig. 3c). *In vivo* interconversion was confirmed by injecting a 1:1 mixture of $GFP^{+}/HER2^{+}$ and $GFP^{-}/HER2^{-}$ CTCs (or the converse), followed by dual GFP and HER2 IHC. Within mixed tumours, GFP-tagged $HER2^{-}$ CTCs produced $GFP^{+}/HER2^{+}$ cells (44%), and in separate tumours, GFP-tagged $HER2^{+}$ CTCs generated $GFP^{+}/HER2^{-}$ cells (21%) (Fig. 2f and Extended Data Fig. 3d).

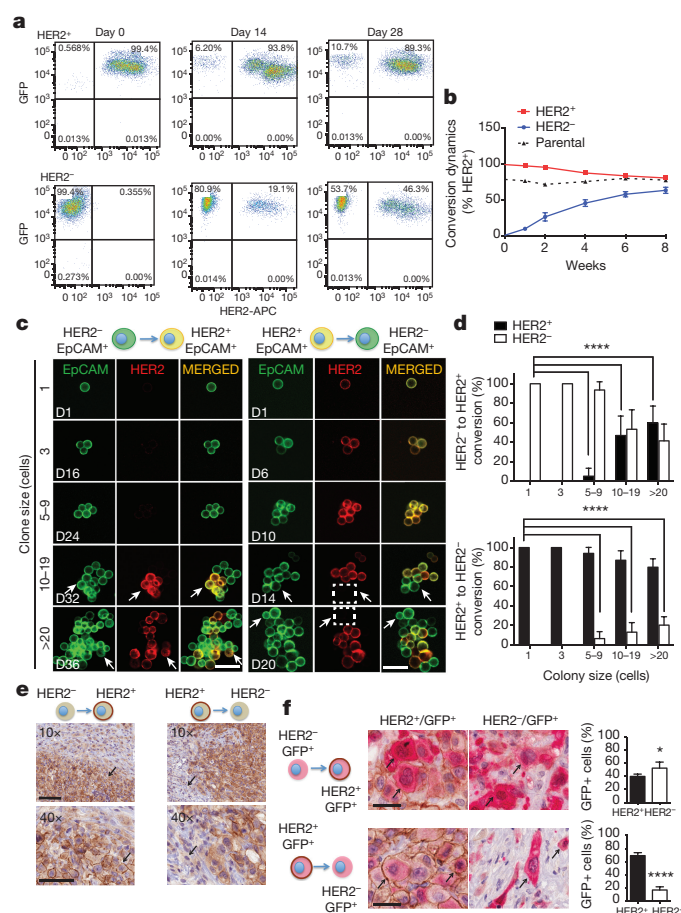


Figure 2 | Interconversion of $HER2^{+}$ and $HER2^{-}$ phenotypes. **a**, FACS-purified GFP -tagged $HER2^{+}$ and $HER2^{-}$ CTCs generate $HER2^{-}$ (top) and $HER2^{+}$ cells (bottom), respectively. **b**, Time course of $HER2^{+}/HER2^{-}$ interconversion following FACS-isolation of $HER2^{+}$ (red) and $HER2^{-}$ (blue) cells; $n = 3$; s.d. (error bar). Parental cultured CTCs (black dotted) are shown as control. **c**, Representative confocal microscopic images depicting $HER2^{+}/HER2^{-}$ interconversion within single-cell-derived clones at indicated time points (D, days) and colony sizes. EpCAM (green), HER2 (red) and MERGED (gold). Scale bar, 20 μ m; $n = 20$. Arrows and dashed boxes indicate interconverting cells, with loss/gain of HER2. **d**, Quantitation of $HER2^{+}/HER2^{-}$ interconversion from single-cell-derived colonies at each colony size; t -test, $P < 0.0001$; $n = 20$; s.d. (error bar). **e**, Gain/loss of $HER2^{+}$ cells (brown, arrow) in tumour xenografts derived from purified $HER2^{-}$ (left)/ $HER2^{+}$ (right) CTCs. Scale bar, 100 μ m (top); 50 μ m (bottom); $n = 8$. **f**, IHC imaging and quantitation of $GFP^{+}/HER2^{+}$ cells within tumours generated from GFP -tagged $HER2^{-}$ and untagged $HER2^{+}$ CTCs (top), and the converse (bottom). GFP : cytoplasmic red; HER2: membrane brown. Scale bar, 20 μ m; t -test, $*P < 0.05$, $****P < 0.0001$; $n = 6$; s.d. (error bar).

To define the molecular characteristics of $HER2^{+}$ versus $HER2^{-}$ CTCs, we quantitatively mapped the global proteomes (>6,300 proteins) of FACS-purified subpopulations (Brx-42, Brx-82, Brx-142) using multiplexed mass spectrometry (MS) with isobaric tandem mass tags (TMT)⁵ (Supplementary Table 3). While proteome profiles of individual cell lines were distinct, they shared differences between $HER2^{+}$ and $HER2^{-}$ subpopulations (Pearson correlation coefficients: Brx-82 versus Brx-142 = 0.81; Brx-82 versus Brx-42 = 0.71; Brx-42 versus Brx-142 = 0.64) (Fig. 3a and Extended Data Fig. 4a, b). $HER2^{+}$ CTCs showed enrichment (Pathway Interaction Database (PID)) of receptor tyrosine kinase (RTK) and pro-growth signalling (GSEA, false discovery rate (FDR) ≤ 0.25) (Fig. 3b and Supplementary Tables 3 and 4). Phosphotyrosine blots from the $HER2^{+}$ subpopulations confirmed RTK phosphorylation (HER2, HER3, HER4, insulin receptor (INSR), EPHA1, EPHA2 and EPHA10), which was absent from matched

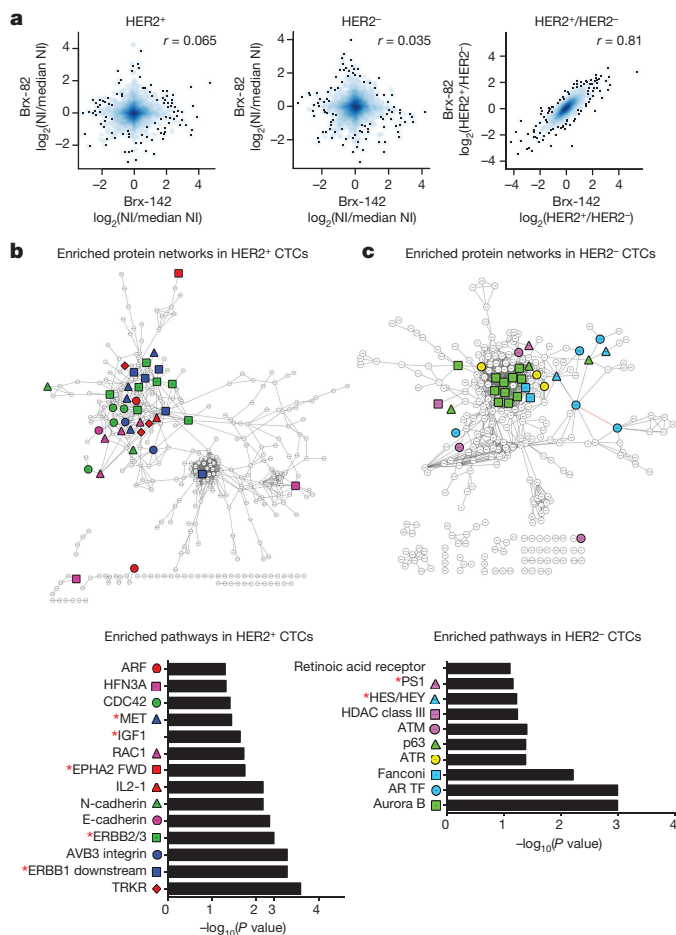


Figure 3 | Molecular pathways differentially activated in HER2⁻ versus HER2⁺ cultured CTCs. **a**, Comparison of quantitative MS proteomes (6,349 proteins) showing distinct profiles for individual cultured CTCs (Brx-82, Brx-142), but linear correlation between proteins differentially expressed in HER2⁺ and HER2⁻ subpopulations; NI, normalized intensity; $n = 2$ biological replicates per CTC line Brx-42, Brx-82, Brx-142 (Supplementary Table 3). **b**, **c**, Cytoscape network maps (top) and GSEA pathway analysis (bottom) depicting proteins enriched by greater than $\log_2(0.5)$ by quantitative MS in (b) HER2⁺ and (c) HER2⁻ CTCs (GSEA FDR ≤ 0.25 ; nominal P cut-off < 0.05 ; Supplementary Table 4). Coloured shapes represent proteins within denoted pathways. Red asterisks highlight RTK pathways in **b**, and Notch pathways in **c**.

HER2⁻ CTCs (Extended Data Fig. 4c). scRNA-seq analysis of 15 primary HER2⁺ CTCs compared with 7 HER2⁻ CTCs from matched patient blood samples showed enrichment for 15 of 32 shared pathways (ERBB1, ERBB2/ERBB3, IGF1, EPHA2, MET) identified by MS analysis of cultured CTC lines (Fig. 3b, Extended Data Fig. 4d, e and Supplementary Tables 4 and 5). In contrast to HER2⁺ CTCs, MS analysis of cultured HER2⁻ CTCs showed increased expression of proteins enriched in Notch (HES/HEY, Presenilin 1 (PS1)) and DNA damage pathways (AuroraB, ATM, ATR, Fanconi) (GSEA, FDR ≤ 0.25) (Fig. 3c and Supplementary Tables 3 and 4).

To explore the potential therapeutic significance of pathways differentially activated in HER2⁺ versus HER2⁻ CTC subpopulations, we screened a panel of 55 drugs selected both for clinical relevance and for the ability to target MS-identified pathways (Supplementary Table 6). HER2⁺ CTCs were no more sensitive to the HER2 inhibitor lapatinib than HER2⁻ CTCs (half-maximum inhibitory concentration (IC₅₀) = 1 μ M), indicating they were not 'oncogene addicted' to HER2, unlike the HER2-amplified SKBR3 cells (IC₅₀ = 5 nM) (Fig. 4a, Extended Data Fig. 5a, b). However, dual inhibition of HER2 and IGF1R, another RTK activated in HER2⁺ CTCs, was cytotoxic to

HER2⁺ but not HER2⁻ CTCs (Fig. 4a), suggesting inhibition of multiple receptor tyrosine kinases may be effective in treating HER2⁺ CTCs. Compared with HER2⁺ CTCs, HER2⁻ CTCs showed reduced sensitivity to the chemotherapeutic agents docetaxel, doxorubicin and 5-fluorouracil (5-FU) (Fig. 4b and Extended Data Fig. 5a, c), but increased sensitivity to γ -secretase inhibitors, which suppress Notch activity (Fig. 4b and Extended Data Fig. 5a, d). Despite proteomic enrichment for Aurora B signalling, HER2⁻ CTCs were not differentially sensitive to Aurora family inhibitors (Fig. 3c, Extended Data Fig. 5a and Supplementary Tables 3 and 4).

The increased NOTCH1 in HER2⁻ CTCs observed by quantitative MS and confirmed by western blot (Extended Data Fig. 6a and Supplementary Tables 3 and 4) was inversely correlated with HER2 expression within primary CTCs and CTC lines, shown by scRNA-seq and immunostaining (Extended Data Fig. 6a). We therefore tested the consequences of suppressing HER2 or activating Notch signalling in HER2⁺ CTCs.

Manipulation of NOTCH1 or its downstream effector NFE2L2/NRF2 in cultured HER2⁺ CTCs did not reduce HER2 expression (Extended Data Fig. 6b). However, inhibition of HER2 using lapatinib or short interfering RNA (siRNA) led to increased expression of NOTCH1, its ligands JAG1 and DLL1, and Notch-regulated genes HES1, HEY1 and HEY2 (Fig. 4c and Extended Data Fig. 6c), confirming previous reports from HER2-amplified breast cancer cells^{6,7} (Extended Data Fig. 6c). Suppression of HER2 also resulted in increased expression of genes (GCLC, GGT1, GPX1, GPX4, HMOX1) downstream of Notch-regulated NRF2, a transcriptional regulator of anti-oxidant/glutathione metabolism pathways^{8,9} (Extended Data Fig. 6d). Thus, expression of HER2 in CTCs appears to mediate downregulation of the NOTCH1/NRF2 axis, potentially switching between proliferative and survival-prone phenotypes.

In addition to suppressing HER2 directly, we tested additional stimuli capable of modulating the HER2⁺/HER2⁻ interconversion. Treatment of HER2⁺ CTCs with low doses of docetaxel (1 nM) or induction of oxidative stress with hydrogen peroxide (H₂O₂; 10 mM) induced rapid shifts from HER2⁺ to HER2⁻ (30% conversion, >70% survival) (Fig. 4d). To exclude differential cell death, we demonstrated acceleration in the appearance of HER2⁻ progeny from FACS-purified single HER2⁺ CTCs (5- to 9-cell stage: 45%; >10-cell stage: 62%) (Fig. 4e and Extended Data Fig. 6e). Thus, exposure to cytotoxic/oxidative stress mediates a switch to a less proliferative but more drug-resistant phenotype.

To model the potential significance of HER2⁺/HER2⁻ interconversion *in vivo*, we generated orthotopic mammary xenografts from FACS-purified subpopulations and analysed tumours before and after treatment with paclitaxel. Purified HER2⁺ CTCs generated mixed tumours (88% HER2⁺, 12% HER2⁻) and showed dramatic tumour shrinkage following paclitaxel treatment. The recurrent tumour showed a transient reduction in HER2⁺ with a corresponding increase in HER2⁻ composition following chemotherapy (2 weeks: 39% HER2⁺; 7 weeks: 74% HER2⁺; Fig. 4f). Purified HER2⁻ CTCs also gave rise to a mixed tumour (35% HER2⁺, 65% HER2⁻), but paclitaxel induced only a limited delay in tumour growth with a minimal effect on HER2 content. Shedding of CTCs was also suppressed by paclitaxel in HER2⁺ but not HER2⁻ tumours (Extended Data Fig. 6f). The chemotherapy-induced shift in HER2 composition was also evident following inoculation of parental CTC cultures (untreated 65% HER2⁺; post-therapy 30% HER2⁺; Extended Data Fig. 6g). Finally, we generated tumours from a 1:1 mixture of GFP-tagged HER2⁺ and untagged HER2⁻ cells, demonstrating a shift from GFP⁺/HER2⁺ to GFP⁺/HER2⁻ cells following paclitaxel treatment (untreated: 70% GFP⁺/HER2⁺, post-therapy: 42% GFP⁺/HER2⁺; Extended Data Fig. 6h). The potent effect of chemotherapy on HER2⁺/HER2⁻ phenotypes *in vivo* may reflect both reduced drug-sensitivity of HER2⁻ cells, as well as stress-induced HER2⁺ to HER2⁻ switching.

Given the demonstrated susceptibility of HER2⁻ CTCs to Notch inhibitors, we combined paclitaxel with either of two γ -secretase

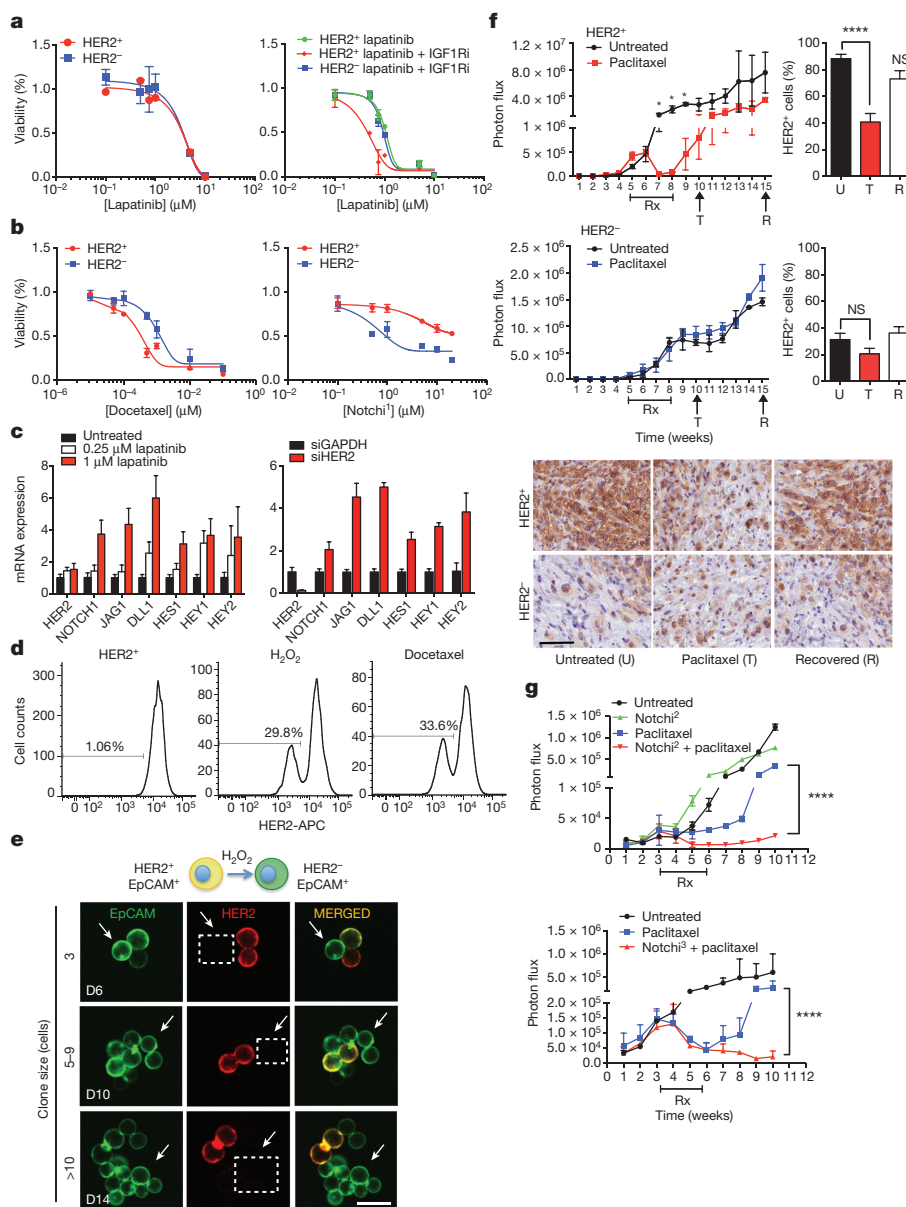


Figure 4 | Cooperative targeting of HER2⁺ and HER2⁻ CTC subpopulations suppresses tumour growth. **a**, HER2⁺ CTCs show no change in sensitivity to lapatinib alone, compared with matched HER2⁻ CTCs (Brx-142), but have increased sensitivity to combined HER2 and IGF1R (BMS-754807) inhibitors; $n = 6$; s.d. (error bar). **b**, HER2⁻ CTCs demonstrate reduced chemosensitivity (docetaxel) but have enhanced sensitivity to Notch inhibition (BMS-708163, Notchi¹), compared with HER2⁺ CTCs; $n = 6$; s.d. (error bar). **c**, Inhibition of HER2 with lapatinib or siRNA-mediated knockdown in HER2⁺ CTCs (Brx-82) results in dose-dependent increase of Notch-related genes: *NOTCH1*, *JAG1*, *DLL1*, *HES1*, *HEY1*, *HEY2*; P (t -test) < 0.05 ; $n = 6$; s.e.m. (error bar). **d**, Rapid emergence (96 h) of HER2⁻ CTCs following treatment of HER2⁺ CTCs with H₂O₂ (10 mM) or docetaxel (1 nM). **e**, Confocal microscopy showing rapid appearance of HER2⁻ progeny from single-CTC derived HER2⁺

inhibitors (LY-411575; RO4929097) in treating mice with tumours initiated from parental CTC lines. Compared with paclitaxel alone, the combination therapy significantly delayed onset of tumour recurrence, while Notch inhibition alone had no effect on tumour growth (Fig. 4g and Extended Data Fig. 6i).

Taken together, we have used primary and cultured CTCs from patients with ER⁺/HER2⁻ breast cancer who developed metastatic multidrug-resistant disease to show that coexisting distinct HER2⁺ and HER2⁻ tumour cell subpopulations may interconvert, with striking

consequences for disease progression and drug response. The comparable tumour initiating potential and similar expression of stem cell marker *ALDH1* in HER2⁺ and HER2⁻ CTCs suggest underlying tumour cell plasticity in these advanced patient-derived breast CTC lines, rather than a hierarchical cancer stem-cell model as described in drug-resistant subpopulations within established breast cancer cell lines^{7,10–15}. While expression of *NOTCH1* and other embryonic markers has been reported in rare, quiescent cells within primary breast tumours^{7,16–18}, the NOTCH1⁺ CTCs reported here constitute a major

cell population, exhibiting both persistent cell proliferation *in vitro* and tumorigenesis *in vivo*. Thus, we propose a dynamic model, in which the equilibrium between HER2⁺ and HER2[−] cells within a heterogeneous tumour population is driven by spontaneous interconversion between these phenotypes, with the more rapidly proliferating HER2⁺ cells prevalent under baseline conditions, and environmental or therapy-induced stress enhancing conversion to the more resistant HER2[−] phenotype. Neither molecular profiling nor functional studies have revealed secreted factors that affect the mutual survival of HER2⁺ and HER2[−] CTCs, but we cannot exclude such additional factors.

Finally, the properties of patient-derived CTC lines established after multiple courses of therapy provide relevant insight to the treatment of drug-refractory, advanced breast cancer. While clinical trials are evaluating the efficacy of HER2-targeted therapy in HER2[−] breast cancer with acquired HER2⁺ CTCs^{1,19–21}, our observations indicate that acquisition of HER2 does not indicate HER2 oncogene dependence and drug susceptibility; instead it constitutes a marker of a proliferative, multi-RTK state. Furthermore, the interconversion of chemotherapy-sensitive HER2⁺/NOTCH1[−] and NOTCH inhibitor-sensitive HER2[−]/NOTCH1⁺ CTCs suggests that dual treatment, as modelled here, may be required for effective treatment. Clinical trials so far have had limited success sequentially administering embryonic pathway inhibitors targeting Hedgehog, Wnt or Notch to inhibit cancer stem cells following initial chemotherapy^{16,22–25}. The rapid interconversion between proliferative and drug-resistant CTC subpopulations raises the possibility that simultaneous combination therapy may provide a novel strategy for clinical validation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 8 December 2015; accepted 25 July 2016.

Published online 24 August 2016.

- Artega, C. L. & Engelman, J. A. ERBB receptors: from oncogene discovery to basic science to mechanism-based cancer therapeutics. *Cancer Cell* **25**, 282–303 (2014).
- Houssami, N., Macaskill, P., Balleine, R. L., Bilous, M. & Pegram, M. D. HER2 discordance between primary breast cancer and its paired metastasis: tumor biology or test artefact? Insights through meta-analysis. *Breast Cancer Res. Treat.* **129**, 659–674 (2011).
- Ozkumur, E. *et al.* Inertial focusing for tumor antigen-dependent and -independent sorting of rare circulating tumor cells. *Sci. Transl. Med.* **5**, 179ra47 (2013).
- Yu, M. *et al.* Ex vivo culture of circulating breast tumor cells for individualized testing of drug susceptibility. *Science* **345**, 216–220 (2014).
- Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature Methods* **8**, 937–940 (2011).
- Osipo, C. *et al.* ErbB-2 inhibition activates Notch-1 and sensitizes breast cancer cells to a γ -secretase inhibitor. *Oncogene* **27**, 5019–5032 (2008).
- Abravanel, D. L. *et al.* Notch promotes recurrence of dormant tumor cells following HER2/neu-targeted therapy. *J. Clin. Invest.* **125**, 2484–2496 (2015).
- DeNicola, G. M. *et al.* Oncogene-induced Nrf2 transcription promotes ROS detoxification and tumorigenesis. *Nature* **475**, 106–109 (2011).
- Wakabayashi, N. *et al.* Notch-Nrf2 axis: regulation of Nrf2 gene expression and cytoprotection by notch signaling. *Mol. Cell. Biol.* **34**, 653–663 (2014).
- Korkaya, H. & Wicha, M. S. HER-2, notch, and breast cancer stem cells: targeting an axis of evil. *Clin. Cancer Res.* **15**, 1845–1847 (2009).
- Ithimakin, S. *et al.* HER2 drives luminal breast cancer stem cells in the absence of HER2 amplification: implications for efficacy of adjuvant trastuzumab. *Cancer Res.* **73**, 1635–1646 (2013).
- Korkaya, H., Paulson, A., Iovino, F. & Wicha, M. S. HER2 regulates the mammary stem/progenitor cell population driving tumorigenesis and invasion. *Oncogene* **27**, 6120–6130 (2008).
- Ginestier, C. *et al.* ALDH1 is a marker of normal and malignant human mammary stem cells and a predictor of poor clinical outcome. *Cell Stem Cell* **1**, 555–567 (2007).
- Martz, C. A. *et al.* Systematic identification of signaling pathways with potential to confer anticancer drug resistance. *Sci. Signal.* **7**, ra121 (2014).
- Pandya, K. *et al.* Targeting both Notch and ErbB-2 signalling pathways is required for prevention of ErbB-2-positive breast tumour recurrence. *Br. J. Cancer* **105**, 796–806 (2011).
- Takebe, N., Harris, P. J., Warren, R. Q. & Ivy, S. P. Targeting cancer stem cells by inhibiting Wnt, Notch, and Hedgehog pathways. *Nature Rev. Clin. Oncol.* **8**, 97–106 (2011).
- Vanharanta, S. & Massagué, J. Origins of metastatic traits. *Cancer Cell* **24**, 410–421 (2013).
- Lawson, D. A. *et al.* Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature* **526**, 131–135 (2015).
- Bidard, F. C. & Pierga, J. Y. Clinical utility of circulating tumor cells in metastatic breast cancer. *J. Clin. Oncol.* **33**, 1622 (2015).
- Schramm, A. *et al.* The DETECT Study Program: personalized treatment in advanced breast cancer based on circulating tumor cells (CTCs). *ASCO Meet. Abstr.* **33**, TPS11109 (2015).
- Ignatiadis, M. *et al.* Abstract OT1-2-02: trastuzumab in HER2-negative early breast cancer as adjuvant treatment for circulating tumor cells (CTCs) (Treat CTC). *Cancer Res.* **75**, OT1-2-02 (2015).
- Amakye, D., Jagani, Z. & Dorsch, M. Unraveling the therapeutic potential of the Hedgehog pathway in cancer. *Nature Med.* **19**, 1410–1422 (2013).
- Kim, E. J. *et al.* Pilot clinical trial of hedgehog pathway inhibitor GDC-0449 (vismodegib) in combination with gemcitabine in patients with metastatic pancreatic adenocarcinoma. *Clin. Cancer Res.* **20**, 5937–5945 (2014).
- LoRusso, P. M. *et al.* Phase I trial of hedgehog pathway inhibitor vismodegib (GDC-0449) in patients with refractory, locally advanced or metastatic solid tumors. *Clin. Cancer Res.* **17**, 2502–2511 (2011).
- Krop, I. *et al.* Phase I pharmacologic and pharmacodynamic study of the gamma secretase (Notch) inhibitor MK-0752 in adult patients with advanced solid tumors. *J. Clin. Oncol.* **30**, 2307–2313 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the patients who participated in this study. This work was supported by National Institutes of Health (NIH) 2R01CA129933, the Howard Hughes Medical Institute, the Breast Cancer Research Foundation, the National Foundation for Cancer Research (DAH) and Wellcome Trust 102696 (C.B.), NIH Quantum 2U01EB012493 (M.T., D.A.H.), T32 CA009361, Susan G. Komen Foundation PDF16376429 (N.V.J.), K12 5K12CA087723 (A.B.) and T32GM007753 (R.Y.E.). We thank D. Dombrowski (NIH 1S100D1016372-01) for expert flow cytometry.

Author Contributions N.V.J., D.A.H. and S.M. conceived the project and provided project leadership. A.B. enrolled patients and provided clinical guidance. B.S.W. and S.R. performed the bioinformatics analyses. M.L. and Y.Z. assisted with animal experiments. T.K.S., M.L.O. and A.J.I. performed the mutational analysis and fluorescence *in situ* hybridization. J.A.L., R.D., R.O. and R.Y.E. picked micromanipulated CTCs for scRNA-seq and assisted with molecular biology experiments. D.S. analysed pathology specimens. C.B. and M.Y. helped with drug screens. S.R. and D.T.T. provided scRNA-seq support. W.H. and M.B. performed MS experiments and analysis. R.V. and M.T. collaboratively developed the CTC-iChip isolation of viable CTCs.

Author Information Single-cell RNA-seq data have been deposited in the Gene Expression Omnibus under accession number GSE75367. Mass spectrometry raw data have been deposited in the MassIVE proteomics data repository under accession number MSV000079419 (<https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp>). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.A.H. (dhaber@mgh.harvard.edu) or S.M. (maheswaran@helix.mgh.harvard.edu).

Reviewer Information Nature thanks J. P. Medema and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Patient selection and CTC isolation. Patients with a diagnosis of metastatic breast cancer provided informed consent for de-identified blood collection, as per institutional review board approved protocol (DF/HCC 05-300). Enrolled patients had received multiple courses of therapy, which is typical in advanced ER⁺ breast cancer, and we did not have sufficient power in this pilot study to enable a statistically significant correlation between the number of therapeutic interventions and the frequency of HER2⁺ CTCs. Patient-matched primary and metastatic tumour specimens were collected according to institutional review board approved protocol (2002-P-002059), and relevant tumour source data are provided in Supplementary Table 1.

Single CTCs were isolated from fresh whole blood by depleting leukocytes using the microfluidic CTC-iChip as previously described³. Briefly, whole blood samples were incubated with biotinylated antibodies against CD45 (R&D Systems, clone 2D1), CD66b (AbD Serotec, clone 80H3) and CD16 (BD, clone 3G8) followed by incubation with Dynabeads MyOne Streptavidin T1 (Invitrogen) to achieve magnetic labelling of white blood cells. This mixture was processed through the CTC-iChip, and the CTCs were stained in solution with Alexa 488-conjugated antibodies against EpCAM (Cell Signaling Technology, clone VU1D9) and HER2 (Cell Signaling Technology, clone 29D8 or Janssen R&D) and identified by imaging flow cytometry (Amnis). Individual CTCs were picked after staining as described above, and PE-CF594-conjugated antibody against CD45 (BD Biosciences, clone HI30) was included to stain contaminating leukocytes. CTCs were individually micromanipulated using a 10 µm transfer tip on an Eppendorf TransferMan NK 2 micromanipulator, transferred into PCR tubes containing RNA protective lysis buffer, and flash frozen in liquid nitrogen as previously described²⁶. Standard CTC enumeration of fixed samples is performed on the BioView high content imaging system following Megafunnel fixation and staining with the combination of wide spectrum cytokeratin (Abcam, ab9377), EpCAM (Cell Signaling Technology, clone VU1D9), EGFR (Cell Signaling Technology, clone D38B1) and HER2 (Cell Signaling Technology, clone 29D8) antibodies.

For mouse xenograft studies, blood was collected via cardiac puncture and ~1 ml of blood was processed through the microfluidic CTC iChip. CTCs were enumerated on the BioView imaging system after staining with Alexa 488-conjugated antibodies against EpCAM (Cell Signaling Technology, clone VU1D9), HER2 (Janssen R&D or Cell Signaling Technology, clone 29D8) and GFP (ab13970) followed by secondary antibodies conjugated with Alexa-488 (Invitrogen).

Immunohistochemistry. Tissues were sectioned, and slides were incubated in 0.3% hydrogen peroxide in methanol for 20 min to block endogenous peroxidase activity. Tissues were permeabilized, and antigen retrieval was performed in 1 × citrate buffer (pH 6) for 15 min. Slides were washed and blocked for 30 min with 5% goat serum. Primary HER2 (Cell Signaling, 29D8) or GFP (Living Colours AV 632381) antibodies were diluted 1:75 or 1:250 in DAKO antibody diluent and samples were incubated for 1 h at room temperature. Slides were incubated with HRP anti-rabbit antibody (EnVision + DAKO) for 30 min. After washing with PBS, the peroxidase reaction was performed with 3,3'-diaminobenzidine (DAB) from Vector Laboratories for 10 min. Cells were counterstained with Gill's #2 haematoxylin for 10–15 s, dehydrated with ethanol and cleared with xylene before mounting. Images represent at least five independent fields from six to eight xenograft tumours per condition.

Fluorescence *in situ* hybridization. Fluorescence *in situ* hybridization was performed as described previously^{27,28}. Briefly, 5-µm sections of formalin-fixed, paraffin-embedded tumour samples were de-paraffinized, hydrated and pre-treated with 0.1% pepsin for 1–2 h. Slides were then washed in 2 × saline-sodium citrate buffer (SSC), dehydrated, air dried and co-denatured at 80 °C for 5 min with a mixture of *CEP17* and *HER2* probes and hybridized at 40 °C overnight using the Hybrite Hybridization System (Abbott). Two-minute post-hybridization washes were performed in 2 × SSC/0.3%NP40 at 72 °C followed by a 1 min wash in 2 × SSC at room temperature. Slides were mounted with Vectashield containing 4',6-diamidino-2-phenylindole (Vector, Burlingame, California, USA). Entire sections were observed with an Olympus BX61 fluorescent microscope equipped with a charge-coupled device camera and analysed with Cytovision software (Applied Imaging, Santa Clara, California).

The *HER2* and *CEP17* signals were quantified in 50 randomly selected, non-overlapping nuclei, and mean numbers of *HER2* and *CEP17* copies per nucleus were calculated. *HER2* was considered amplified when the *HER2:CEP17* ratio was ≥2.0 or *HER2* signals per nuclei was >6 following the guidelines of the American Society of Clinical Oncology/College of American Pathologists²⁹. The probes used in this study consisted of centromeric *CEP*: 17p11.1-q11.1, spectrum aqua (Abbott Molecular, Des Plaines, Illinois) and locus-specific identifier probes derived from bacterial artificial chromosome RP11-94L15 (17q12-17q21.1, spectrum orange probe (CHORI, Oakland, California)).

CTC cell culture. CTC cultures were grown in suspension in ultra-low attachment plates (Corning) in tumour sphere medium (RPMI-1640, EGF (20 ng/ml), bFGF (20 ng/ml), 1X B27, 1X antibiotic/antimycotic (Life Technologies)) under hypoxic (4% O₂) conditions. The breast CTC lines, Brx-42, Brx-82 and Brx-142, were derived from CTCs isolated using the CTC-iChip as previously described⁴. CTC lines were routinely checked for mycoplasma, using a mycoplasma detection kit (MycoAlert, Lonza), and were authenticated by RNA-seq, MS and DNA-seq (1,000 gene mutation panel).

Fluorescence-activated cell sorting (FACS). Cells were trypsinized into single-cell suspensions, resuspended in Hanks' balanced salt solution (HBSS), and incubated with Anti-HER2/NEU APC (BD, clone 42 c-erbB-2), Anti-HER2 FITC (Janssen R & D) or Annexin V FITC (BD, clone RUO) antibodies for 20 min at 4 °C. Unbound antibodies were washed from cells using HBSS. For analytical flow, cells were fixed with 3% paraformaldehyde and analysed using a Laser BD Fortessa instrument. For sterile live-cell flow cytometry, cells were sorted using a Laser BD FACS Aria Fusion Cell Sorter, BSL2⁺. FACS plots are representative of at least two independent experiments performed within 6 months of culture initiation (Figs 1d and 2a and Extended Data Figs 1f and 3a).

Sequencing analysis of genomic DNA. Genomic DNA extracted from CTC-derived cell lines was sequenced using a multiplex polymerase chain reaction (PCR) technology called Anchored Multiplex PCR (AMP) for single nucleotide variant (SNV) and insertion/deletion (indel) detection using next generation sequencing (NGS) as previously described³⁰. Briefly, genomic DNA was isolated from cell lines and then sheared with the Covaris M220 instrument, followed by end-repair, adenylation and ligation with an adaptor. A sequencing library targeting hotspots and exons in 39 commonly mutated, cancer-associated genes was generated using two hemi-nested PCR reactions. Illumina MiSeq 2 × 151 base paired-end sequencing results were aligned to the hg19 human genome reference using BWA-MEM³¹. MuTect³² and a laboratory-developed insertion/deletion analysis algorithm were used for SNV and indel variant detection, respectively. This assay has been validated to detect SNV and indel variants at 5% allelic frequency or higher in target regions with sufficient read coverage.

Lentivirus production, infection and siRNA knockdown of CTC cell lines. To produce replication-incompetent lentivirus, 293T cells were co-transfected with either Lenti-Luc-GFP or Notch intracellular domain-pcw107 (Addgene 64621) constructs in combination with REV, VSVG, PDML or pMD2.G and psPAX2 (Addgene) using Lipofectamine Plus reagent (Invitrogen). Twenty-four hours later, growth medium was replenished. Viral supernatants were harvested 48 h post-transfection, concentrated with Lenti-X Concentrator (Clontech), and viral pellets were resuspended in 400 µl base medium. CTC cultures were infected overnight with 100 µl lentivirus in 6 µg/ml Polybrene. Puromycin (3 µg/ml) was used to select transduced cells over a period of 7 days. For the RNAi knockdown, CTC lines Brx-42, Brx-82 and Brx-142 were reverse transfected in ultra-low attachment six-well plates (Corning) with 25 nM siRNA smart pools (Dharmacon) containing the combination of four different siRNA oligonucleotides for *ERBB2/HER2* (GGACGAAUUCUGCACAAG; GACGAAUUCUGCACAAGG; CUACAACAGACAGCGUUU; AGACGAAGCAUACGUGAUG), *NOTCH1* (GCGACAAGGUGUUGACGUG; GAUGCGAGAUCGACGUGCAA; GAACGGGGCUAACAAGAU; GCAAGGACCACUUCAGCGA), *NRE2L2* (GAGAAAGAAUUGCCUGUAA, CCAAAGAGCAGUUCAAUGA, UAAAGUGGUGCUCAGAAU; UGACAGAAGUUGACAAUUA) or the negative control gene GAPDH. siRNA pools for target genes were deconvolved to demonstrate targeted knockdown efficiency (more than two siRNAs per gene).

Immunofluorescence. CTC lines were spun onto poly-L-lysine-functionalized glass slides by Spintrap, fixed with 3% paraformaldehyde, permeabilized with 0.1% Triton X and stained with nuclear 4,6-diamidino-2-phenylindole (DAPI) stain, HER2 (Cell Signaling Technologies, clone 29D8), Ki67 (Zymed), Cleaved Caspase-3 (Cell Signaling Technologies, clone D3E9) and/or NOTCH1 (Cell Signaling Technologies, clone D1E11) antibodies. Secondary antibodies were conjugated to either Alexa Fluor 488 or Alexa Fluor 594 (Life Technologies), and fluorescence was measured using the Nikon 90-I fluorescent microscope. Images are representative of at least three independent images per sample.

Single-cell lineage tracing and confocal microscopy. Single HER2⁺ or HER2⁻ CTCs were flow sorted in 96-well white-walled plates (Corning) using Laser BD FACS Aria Fusion Cell Sorter, BSL2⁺. Single cell, 1-, 3-, 5-, 10- to 20- and > 20-cell clones were analysed for heterogeneity in HER2 expression via staining with antibodies against EpCAM (FITC labelled; Cell Signaling, clone VU1D9) and HER2 (APC labelled, BD, clone 42 c-erbB-2). Imaging and image processing was performed sequentially with the confocal microscope (Zeiss 710 Laser Scanning Confocal) followed by FIJI (Image J). Images are representative of at least 20 independent images per colony size.

Determination of reads-per-million (RPM). Trimmomatic was used to crop reads lengths to 50 nucleotides, and to remove the TruSeq3-PE-2 Illumina adapters. The paired-end reads were then aligned using tophat2 and bowtie1 with the no-novel-juncs argument set with human genome version hg19 and transcriptome defined by the hg19 genes.gtf table from <http://genome.ucsc.edu>. Reads that did not align or aligned to multiple locations were discarded. The number of reads aligning to each gene was then determined using htseq-count. Samples that had fewer than 10^5 reads were discarded. The read count for each gene was divided by the total counts assigned to all genes and multiplied by one million to form the reads per million (RPM). Samples for which the expression of the white blood cell marker PTPRC (CD45) was greater than 10 RPM were discarded. Single-cell RNA-seq data have been deposited in the Gene Expression Omnibus under accession number GSE75367.

Bimodality. To establish that the distribution of HER2 expression in CTCs is multi-modal, we applied the Hartigan's dip test as implemented in the diptest R-package to the $\log_{10}(\text{RPM} + 1)$ values with 10 RPM as the threshold to define HER2⁻ versus HER2⁺ CTCs. To establish that the distribution has two modes and not more, we applied the density function of R with default values to the $\log_{10}(\text{RPM} + 1)$ values.

Gene set enrichment analysis of RNA-seq and quantitative proteomics data. On the basis of the analysis of bimodality above, we defined HER2⁺ samples to be those for which the expression of HER2 exceeded 10 RPM and defined the rest to be HER2⁻. For the mass spectrometric data, enrichment of signalling pathways was determined by submitting the average \log_2 fold-change in protein abundance between the HER2-high and HER2-low samples to the pre-ranked function of the Broad Institute's GSEA software using gene sets in the Pathway Interaction Database (PID) and KEGG as curated in version 4 of the Broad Institute's MSigDB (<http://www.broadinstitute.org/gsea/msigdb/>). Pathway enrichment for the RNA-seq of the CTCs was done the same way with the exception that the full RPM matrix for the CTCs and the HER2⁺ versus HER2⁻ distinction was input to the GSEA software instead of \log_2 fold-change.

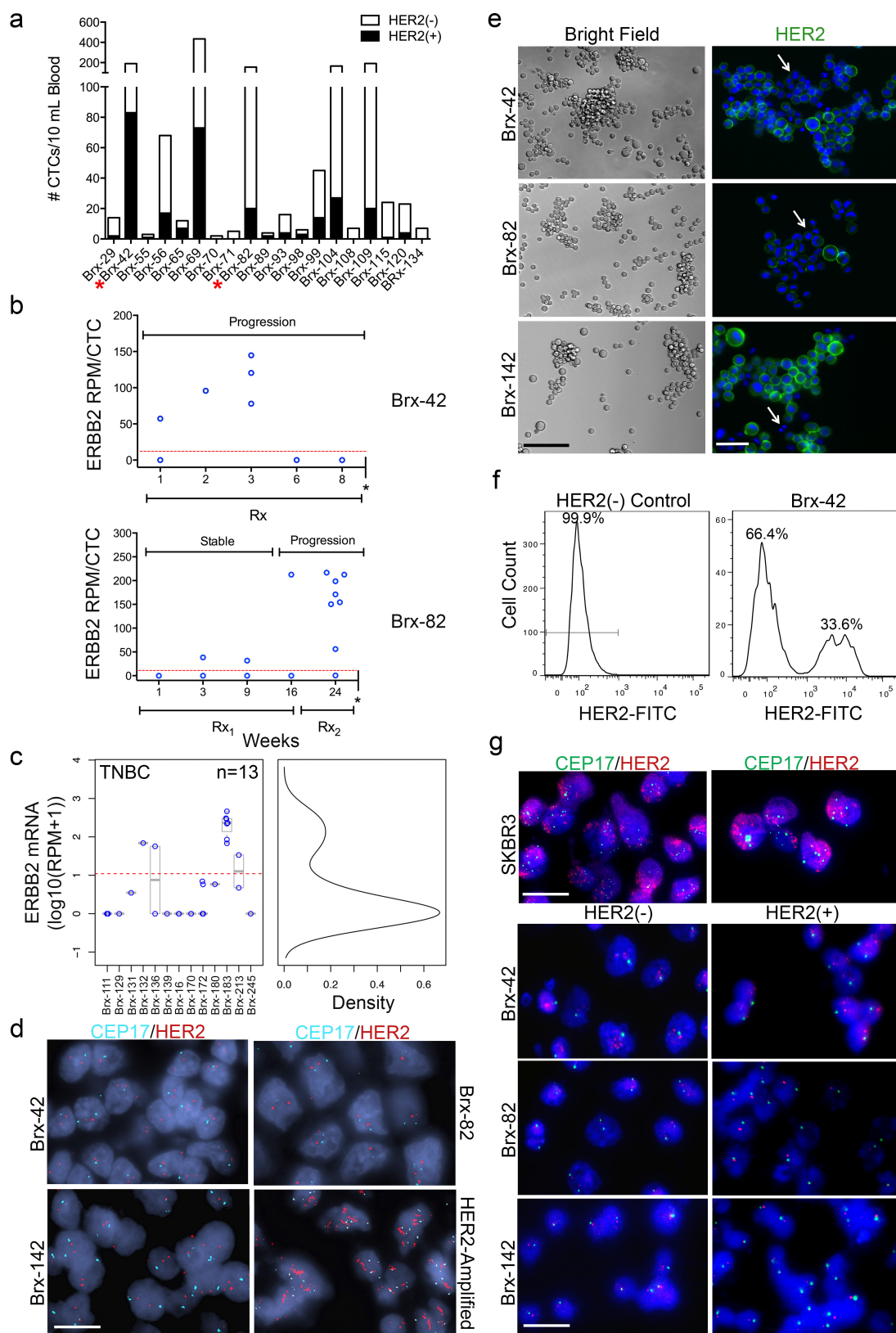
Quantitative proteomics. CTC cell pellets were re-suspended in lysis buffer containing 75 mM NaCl, 50 mM HEPES (pH 8.5), 10 mM sodium pyrophosphate, 10 mM NaF, 10 mM β -glycerophosphate, 10 mM sodium orthovanadate, 10 mM phenylmethanesulfonylfluoride, Roche Complete Protease Inhibitor EDTA-free tablets and 3% sodium dodecyl sulfate. Cells were lysed by passing them ten times through a 21-gauge needle, and the lysates were prepared for analysis on the mass spectrometer essentially as described previously⁵. Briefly, reduction and thiol alkylation were followed by purifying the proteins using MeOH/CHCl₃ precipitation. Protein digest was performed with Lys-C and trypsin, and peptides were labelled with TMT-10plex reagents (Thermo Scientific)³³ and fractionated by basic pH reversed phase chromatography. Multiplexed quantitative proteomics was performed on an Orbitrap Fusion mass spectrometer (Thermo Scientific) using a simultaneous precursor selection (SPS)-based MS3 method³⁴. MS2 spectra were assigned using a SEQUEST-based proteomics analysis platform³⁵. On the basis of the target-decoy database search strategy³⁶ and employing linear discriminant analysis and posterior error histogram sorting, peptide and protein assignments were filtered to a FDR of < 1% (ref. 35). Peptides with sequences that were contained in more than one protein sequence from the UniProt database were assigned to the protein with most matching peptides³⁵. TMT reporter ion intensities were extracted as that of the most intense ion within a 0.03-thomson window around the predicted reporter ion intensities in the collected MS3 spectra. Only MS3 with an average signal-to-noise value larger than 40 per reporter ion as well as with an isolation specificity⁵ larger than 0.75 were considered for quantification. A two-step normalization of the protein TMT-intensities was performed by first normalizing the protein intensities over all acquired TMT channels for each protein on the basis of the median average protein intensity calculated for all proteins. To correct for slight mixing errors of the peptide mixture from each sample, a median of the normalized intensities was calculated from all protein intensities in each TMT channel, and protein intensities were normalized to the median value of these median intensities.

Protein interactions were extracted from the String database (high confidence score > 0.7)³⁷. Overlapping proteins were assigned to the pathway with the greatest number of proteins, and enriched PID pathways were ranked by $\log_{10}(P \text{ value})$ to the nearest thousandth. Mass spectrometry raw data have been deposited in the MassIVE proteomics data repository under the accession number MSV000079419.

Drug screens. Drugs were obtained from the MGH Center for Molecular Therapeutics and are listed in Supplementary Table 6. They were chosen because of their common clinical use for treatment of breast cancer or unique targeting of epigenetic/stem cell pathways. One thousand cells were seeded in tumour sphere media in 384-well ultra-low attachment plates in triplicate wells on duplicate plates 24 h before the addition of drugs. Three independent drug concentrations centred on the reported IC₅₀ were used (Supplementary Table 6). Cell viability was assayed 6 days after drug treatment with CellTiter-Glo (Promega) and was normalized to corresponding untreated controls³⁸.

Mouse xenograft assays and drug treatment. In compliance with ethical regulations and approved by the animal protocol (IACUC 2010N000006), 6-week-old female NSG (NOD. Cg-Prkscsdid Il2rgtm1Wjl/SzJ) mice from Jackson Laboratories were anaesthetized with isoflurane, and GFP-LUC labelled CTCs (200,000, 20,000 and/or limiting dilutions as low as 200 cells) or 50:50 mixed CTCs (GFP-LUC⁺/HER2⁺: Untagged/HER2⁻, and the converse) were injected into the fourth right mammary fat pad. A 90-day release 0.72 mg oestrogen pellet (Innovative Research of America) was implanted subcutaneously behind the neck of each mouse. Tumour growth was monitored weekly by *in vivo* imaging using IVIS Lumina II (PerkinElmer) following intraperitoneal injection (150 μ l per animal) of D-luciferin substrate (Sigma). For *in vivo* drug sensitivity testing, Paclitaxel (10 mg/kg) was administered weekly by intravenous injection for 4 consecutive weeks. Notch inhibitors (Notchi² LY-411575 (10 mg/kg) or (Notchi³) RO429097 (10 mg/kg) were administered daily (5 days on/2 days off) via oral gavage in 2% solvent (2% sodium carboxymethyl cellulose) for 4 consecutive weeks. No animal randomization or blinding was used for these mouse studies. All animal studies used six to eight mice per condition to ensure sufficient statistical power.

26. Miyamoto, D. T. *et al.* RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. *Science* **349**, 1351–1356 (2015).
27. Mohapatra, G. *et al.* Glioma test array for use with formalin-fixed, paraffin-embedded tissue: array comparative genomic hybridization correlates with loss of heterozygosity and fluorescence in situ hybridization. *J. Mol. Diagn.* **8**, 268–276 (2006).
28. Snuderl, M. *et al.* Polysomy for chromosomes 1 and 19 predicts earlier recurrence in anaplastic oligodendrogliomas with concurrent 1p/19q loss. *Clin. Cancer Res.* **15**, 6430–6437 (2009).
29. Wolff, A. C. *et al.* American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *J. Clin. Oncol.* **25**, 118–145 (2007).
30. Zheng, Z. *et al.* Anchored multiplex PCR for targeted next-generation sequencing. *Nature Med.* **20**, 1479–1484 (2014).
31. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
32. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnol.* **31**, 213–219 (2013).
33. McAlister, G. C. *et al.* Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal. Chem.* **84**, 7469–7478 (2012).
34. McAlister, G. C. *et al.* MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal. Chem.* **86**, 7150–7158 (2014).
35. Huttlin, E. L. *et al.* A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**, 1174–1189 (2010).
36. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4**, 207–214 (2007).
37. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39**, D561–D568 (2011).
38. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).

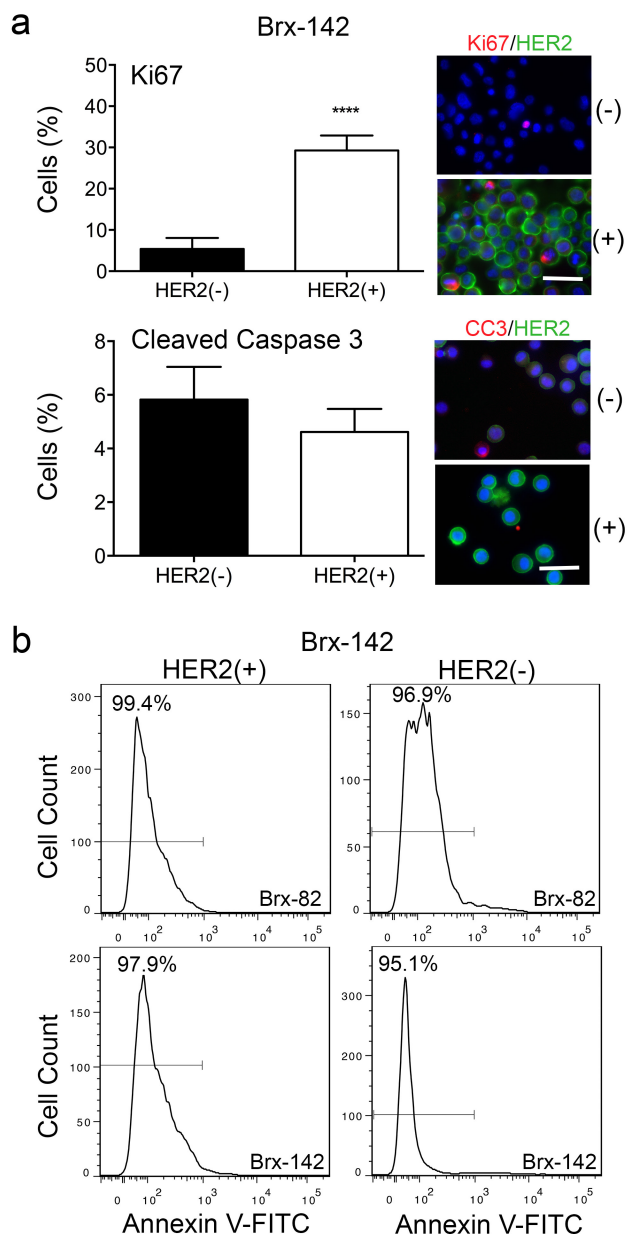


Extended Data Figure 1 | See next page for caption.

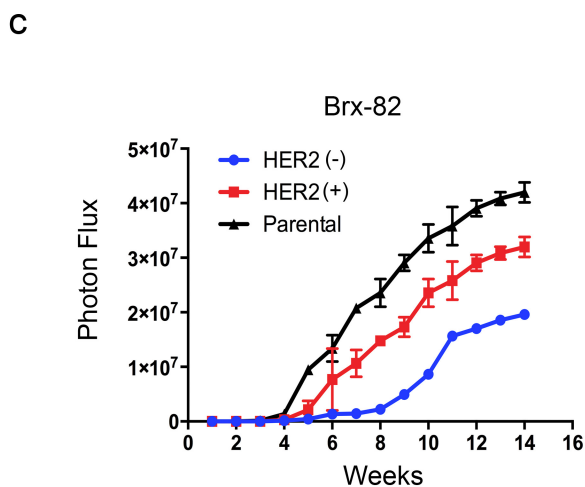
Extended Data Figure 1 | Patients with advanced ER⁺/HER2⁻ breast cancer harbour discrete HER2⁺ and HER2⁻ subpopulations.

a, CTCs freshly isolated from 19 patients with ER⁺/HER2⁻ breast cancer were stained with HER2 (green) and EpCAM (yellow) and imaged using imaging flow cytometry. Bar graph shows the number of HER2⁺ (black) and HER2⁻ (white) CTCs (median 22% HER2⁺ CTCs, range 4–58%). Supplementary Table 1 provides HER2⁺/HER2⁻ ratios and each patient's clinical history. **b**, scRNA-seq for *ERBB2* expression at multiple time-points showing acquisition of HER2⁺ CTCs (Brx-82, Brx-42) over the course of progressive disease. Single asterisk (*) denotes patient expiration. Rx, sacituzumab (IMMU-132); Rx₁, vinorelbine + trastuzumab; Rx₂, eribulin. **c**, Distinct HER2⁺ and HER2⁻ CTCs from 13 patients with triple-negative breast cancer (TNBC) determined by scRNA-seq (HER2⁻ ≤ 0 RPM; HER2⁺ > 153, range 33–463). **d**, HER2 fluorescence *in situ* hybridization (FISH) analysis of metastatic tumours

from patients, Brx-42, Brx-82 and Brx-142, shows no amplification of *ERBB2* compared with HER2-amplified control (Supplementary Table 1 for tumour source data). HER2 (red); chromosome enumeration probe 17 (CEP17) (cyan); scale bar, 10 μm. Representative images from five independent fields are shown. **e**, Bright field and immunofluorescence (DAPI, blue; HER2, green) images of CTC lines, Brx-42, Brx-82 and Brx-142, demonstrate heterogeneity in HER2 expression. Scale bar, 100 μm (bright field); 20 μm (immunofluorescence). Representative images from three independent fields are shown. **f**, FACS analysis shows two distinct HER2⁺ and HER2⁻ subpopulations in the CTC line Brx-42 (at initiation) compared with HER2⁻ control. Representative data of two independent experiments are shown. **g**, HER2 FISH analysis of the HER2⁺ and HER2⁻ subpopulations from CTC lines Brx-42, Brx-82 and Brx-142 shows that *ERBB2* is not amplified. HER2-amplified SKBR3 cells shown as control. HER2 (red); CEP17 (green); scale bar, 10 μm. Representative images from five independent fields are shown.



Extended Data Figure 2 | HER2⁺ and HER2⁻ subpopulations exhibit distinct functional properties. **a**, Increased expression of the proliferation marker Ki67 (red) in the HER2⁺ subpopulation of CTC line Brx-142 (*t*-test, $P < 0.0001$), compared with the HER2⁻ subpopulation, with no change in cleaved-caspase 3 (red). HER2⁺ cells (green); scale bar, 20 μ m. Representative images from five independent fields are shown. **b**, FACS analysis for the apoptotic marker Annexin V-FITC shows no difference in apoptosis between the HER2⁺ and HER2⁻ subpopulations of FACS-



d

Metastatic Frequency from Orthotopic Injections

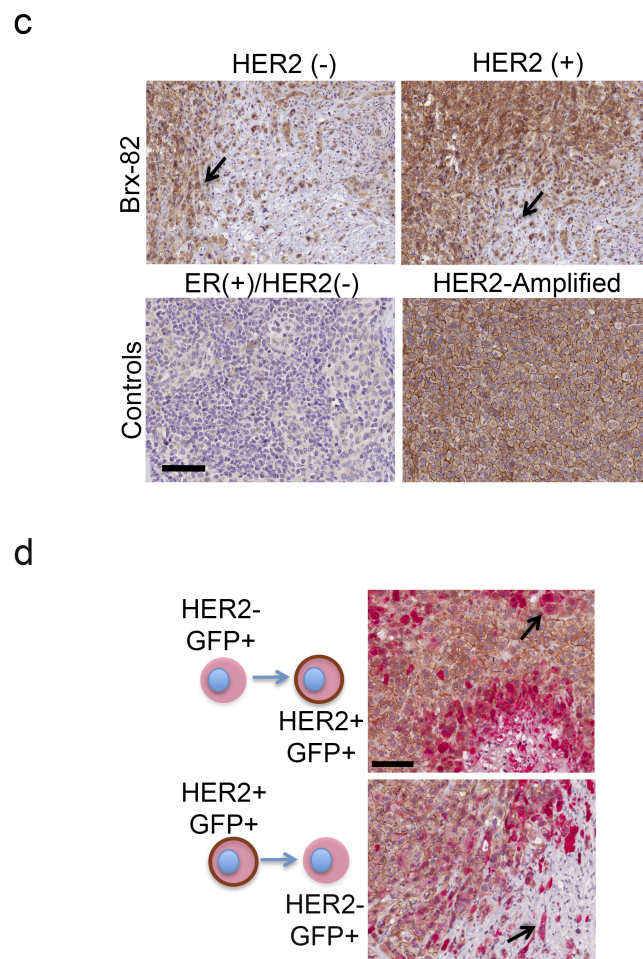
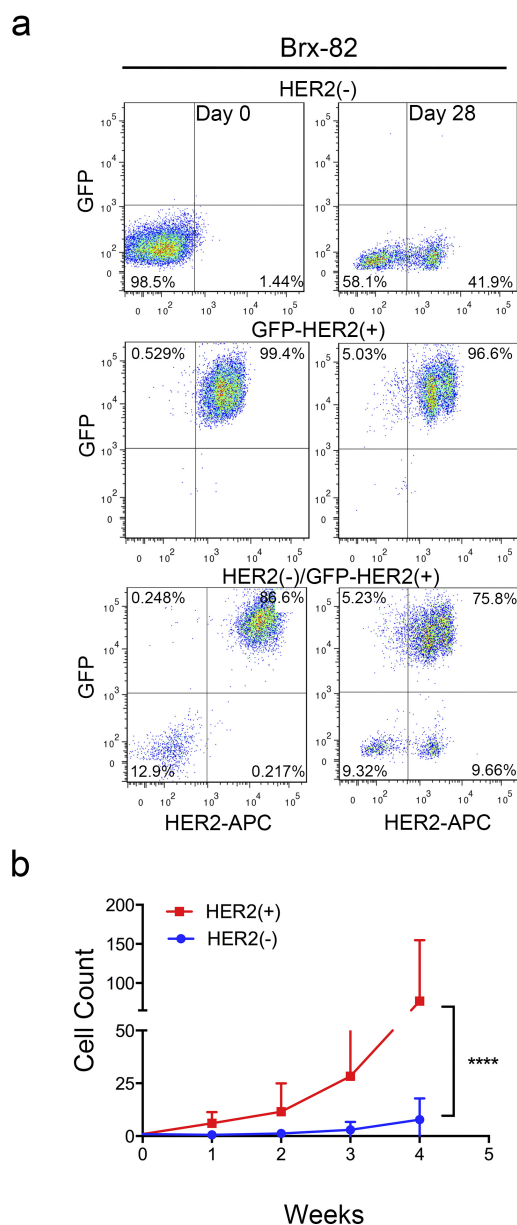
CTC Cell Line	HER2(+)	HER2(-)	P-value
Brx-82	6/8	2/8	0.05
Brx-142	7/8	2/8	0.009

e

Tumor Initiation from 200 Cells

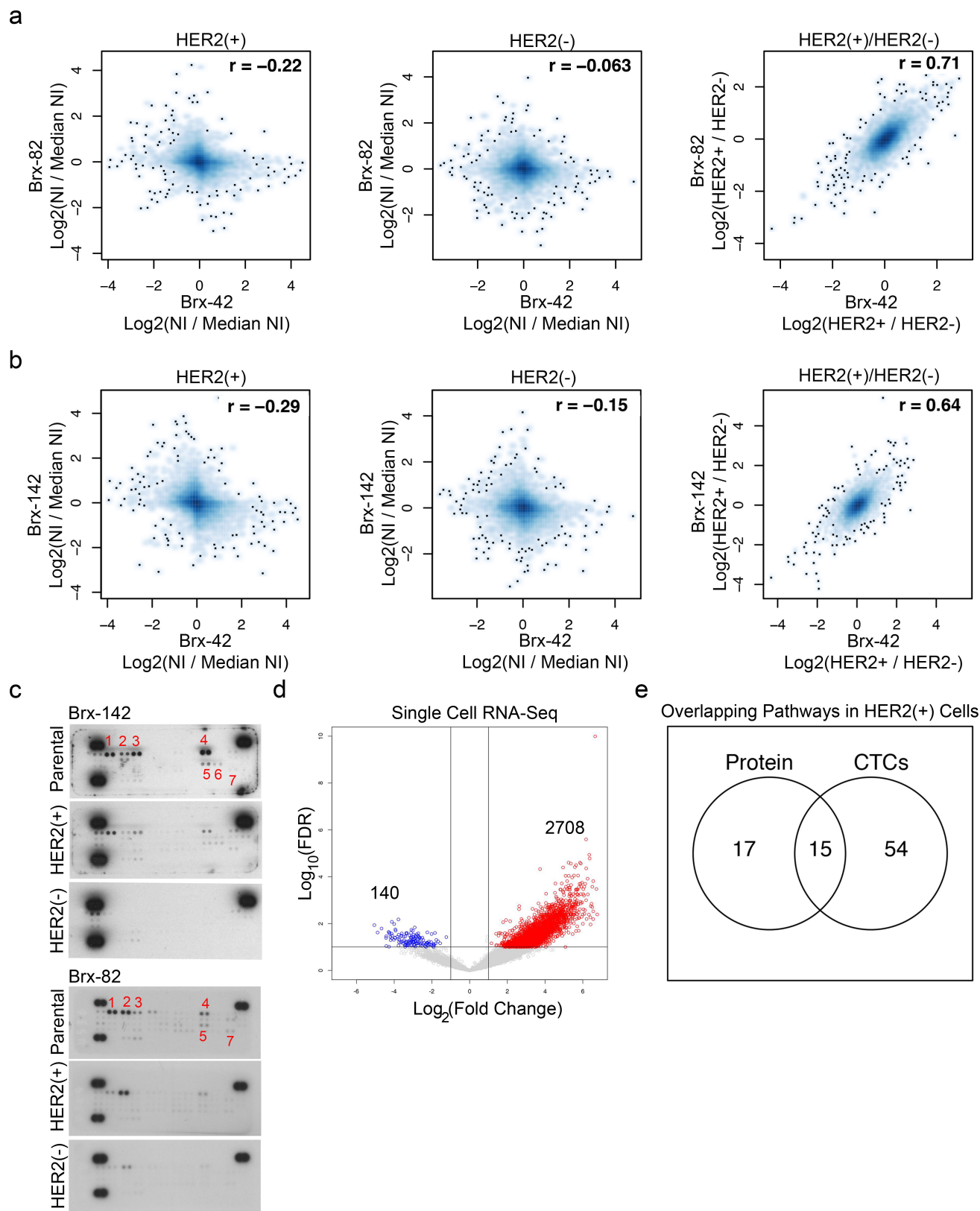
CTC Cell Line	HER2(+)	HER2(-)	P-value
Brx-82	8/8	8/8	NS
Brx-142	4/8	3/8	NS

purified CTC line Brx-142. Representative data from two independent experiments are shown. **c**, Tumours initiated by HER2⁺ or HER2⁻ CTCs (Brx-82: 200,000 cells) orthotopically injected into the mammary fat pad show differential growth rates; $n = 8$. **d**, Metastatic frequency of HER2⁺ and HER2⁻ cultured CTCs (Brx-82: $P = 0.05$; Brx-142: $P = 0.009$) following orthotopic injection; $n = 8$. **e**, Limiting dilution experiments demonstrate comparable tumour initiating ability from 200 HER2⁺ and HER2⁻ cultured CTCs (Brx-82, Brx-142); $n = 8$.



Extended Data Figure 3 | Dynamics of HER2⁺ and HER2⁻ interconversion. **a**, FACS-purified HER2⁺ and HER2⁻ subpopulations from CTC line Brx-82 were monitored over 28 days to determine shifts in the composition of sorted populations. Representative data of two independent experiments are shown. **b**, Growth curves for HER2⁺ (red) and HER2⁻ (blue) FACS-purified single cell clones from CTC line Brx-142; two-way ANOVA, $P < 0.0001$; $n = 20$. **c**, IHC HER2 staining of tumour xenografts derived from unlabelled HER2⁻ and HER2⁺ CTCs showing acquisition/loss of HER2 (brown), respectively. Arrows indicate

regions of HER2 acquisition/loss. Representative image from at least five independent fields; $n = 8$. ER⁺/HER2⁻ and HER2-amplified breast cancers are shown below as controls. **d**, Low-magnification (landscape) view of HER2 IHC staining of tumour xenografts derived from mixed HER2⁺ and HER2⁻ CTC cultures containing either GFP-tagged HER2⁺/HER2⁻ cells (high magnification images are shown in Fig. 2f). Top: representative GFP-tagged HER2⁻ cells give rise to GFP⁺/HER2⁺ cells (GFP: cytoplasmic red stain, HER2: cell surface brown stain). Bottom: GFP-tagged HER2⁺ cells produce GFP⁺/HER2⁻ cells. Scale bar, 100 μm.

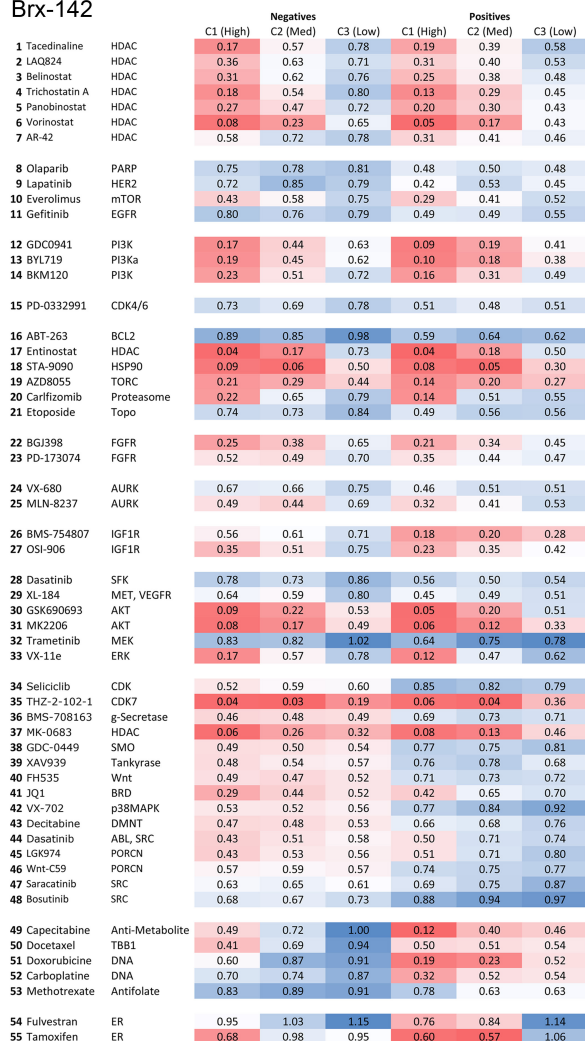


Extended Data Figure 4 | Proteomic and scRNA-seq analysis of HER2⁺ versus HER2⁻ cells. a, b, MS-based whole cell proteome profiles (6,349 proteins) comparing HER2⁺ and HER2⁻ populations from CTC lines (Brx-42, Brx-82, Brx-142). Matched HER2⁺ versus HER2⁻ proteomic differences show significant linear correlation (Pearson correlation coefficient = 0.71 between Brx-82 and Brx-42; Pearson correlation coefficient = 0.64 between Brx-142 and Brx-42); NI, normalized intensity; $n = 2$ per cell line are shown. **c**, Phospho-RTK array of HER2⁺ and HER2⁻ populations of CTC cell lines Brx-142 and Brx-82 show increased

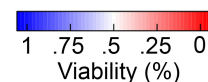
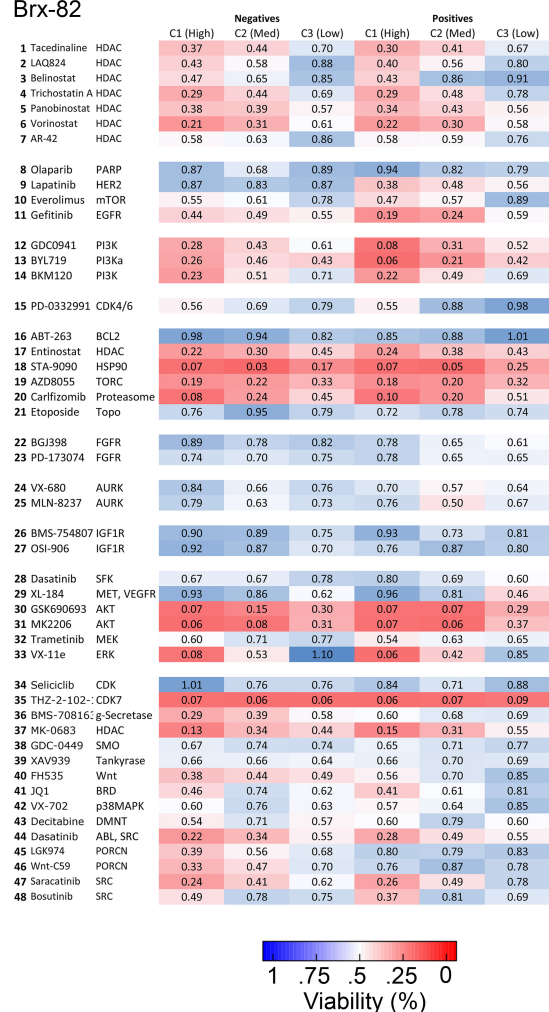
phosphorylation of RTKs in the HER2⁺ population. Numbers denote the following: 1, HER2; 2, HER3; 3, HER4; 4, INSR; 5, EPHA1; 6, EPHA2; 7, EPHA10. Representative data from two independent experiments are shown. **d**, Volcano plot depicts genes enriched in HER2⁺ (red) and HER2⁻ (blue) individual CTCs isolated from patients Brx-42 and Brx-82 and analysed by scRNA-seq; $n = 22$. **e**, Venn diagram showing PID pathway overlap of genes and proteins derived from scRNA-seq (Brx-42, Brx-82) and quantitative proteomics of HER2⁺ CTCs, respectively.

a

Brx-142

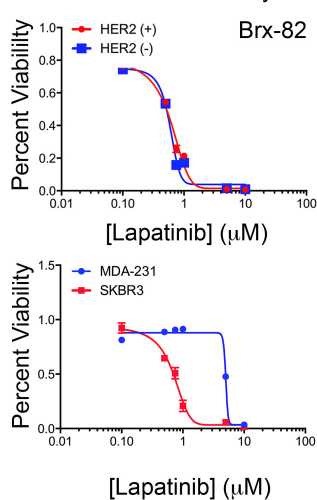


Brx-82



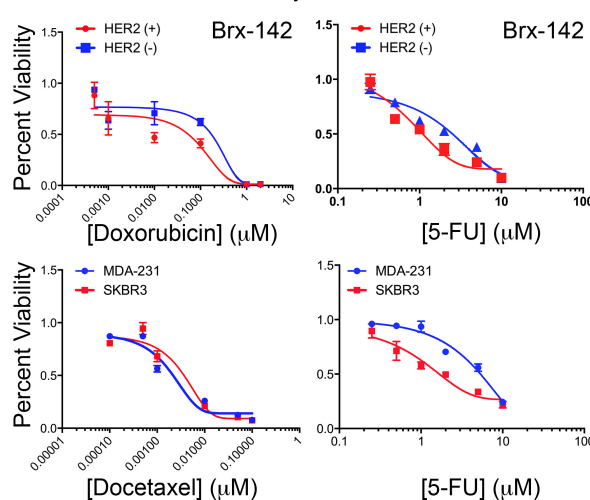
b

HER2 Pathway



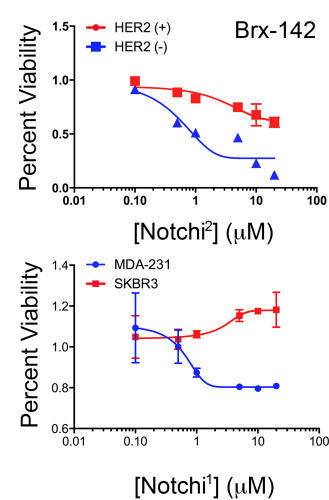
c

Cytotoxic



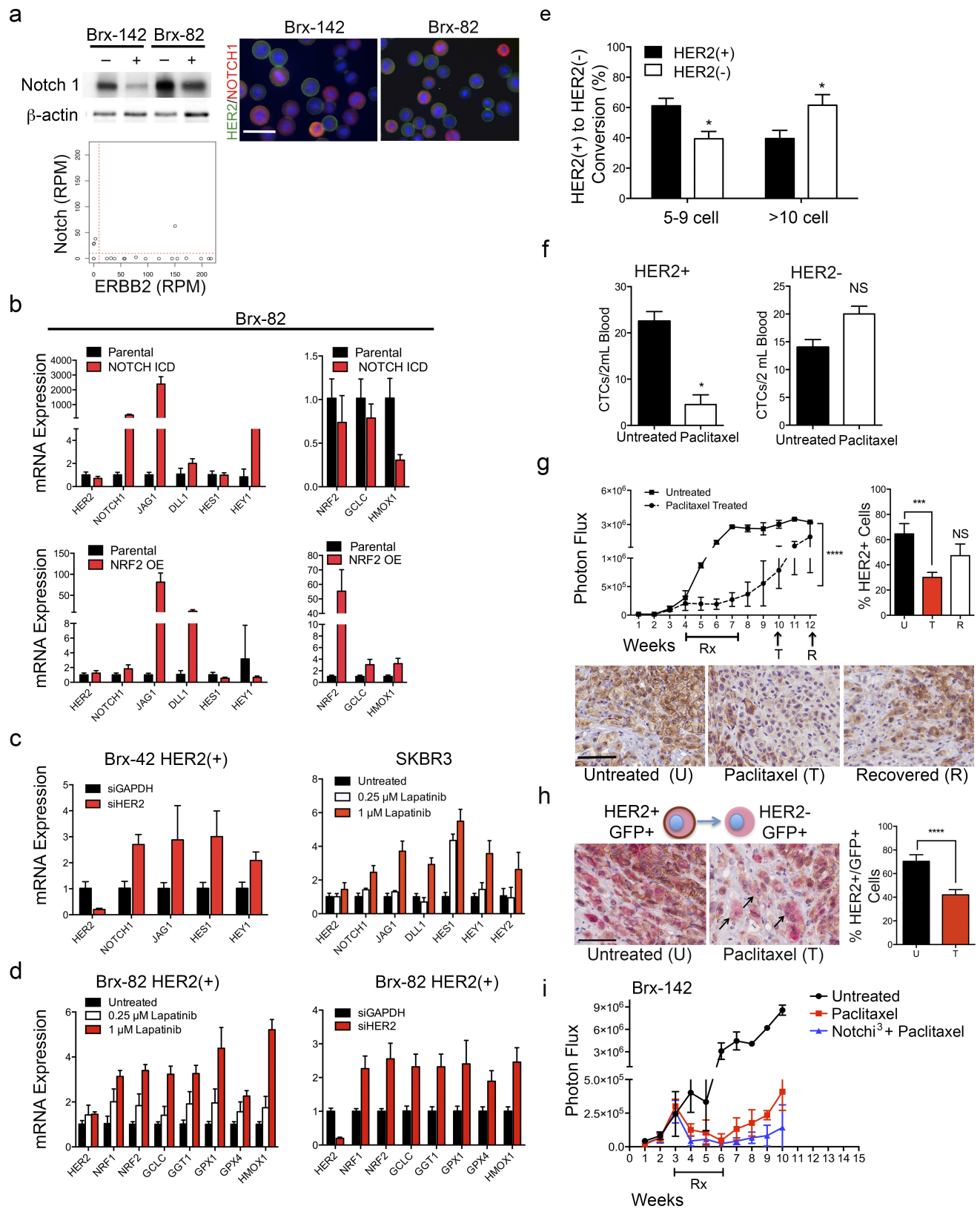
d

Notch Inhibitors

Extended Data Figure 5 | Fifty-five panel drug screen shows differential drug sensitivities exhibited by HER2⁺ versus HER2⁻ subpopulations.

a, Heat map showing percentage cell viability (represented as decimal) after 6 days of drug treatment of the HER2⁺ and HER2⁻ subpopulations derived from CTC lines Brx-142 and Brx-82. Red and blue represent high and low drug sensitivities, respectively; *n* = 6. **b**, Lapatinib sensitivity of HER2⁺ (red) and HER2⁻ (blue) subpopulations of CTC line Brx-82. MDA-231 (TNBC) and SKBR3 (HER2-amplified) are shown as controls.

c, Chemosensitivity of HER2⁺ (red) and HER2⁻ (blue) subpopulations of CTC line Brx-142. MDA-231 (blue) and SKBR3 (red) are shown as controls. **d**, Sensitivity of HER2⁺ (red) and HER2⁻ (blue) subpopulations of CTC line Brx-142 to Notch inhibition with Notch1¹ (BMS-708163) and Notch2² (RO4929097). MDA-231 and SKBR3 cells are shown as controls. **a–d**, Representative of at least two independent experiments for each condition; *n* = 6.



Extended Data Figure 6 | See next page for caption.

Extended Data Figure 6 | NOTCH1 expression and activity in HER2[−] CTCs. **a**, Western blot analysis of HER2⁺ and HER2[−] subpopulations from CTC lines Brx-142 and Brx-82 show increased NOTCH1 in HER2[−] cells. β -Actin is shown as control. Immunofluorescence analysis and scRNA-seq of NOTCH1 (red) and HER2 (green) shows inversely correlated expression in CTC lines (Brx-142, Brx-82). **b**, Ectopic expression of constitutively active Notch intracellular domain (ICD) or NRF2 results in increased expression of the Notch1 ligand *JAG1* but does not alter HER2 expression. Representative data of two independent experiments are shown; s.e.m. (error bars). **c**, siRNA-mediated inhibition of HER2 in Brx-42 HER2⁺ CTCs, and lapatinib-mediated inhibition of HER2 in SKBR3 cells results in dose-dependent increases in the expression of genes involved in Notch signalling (*NOTCH1*, *JAG1*, *DLL1*, *HES1*, *HEY1*, *HEY2*). Representative data of two independent experiments are shown; s.e.m. (error bars). **d**, Inhibition of HER2 using lapatinib or siRNA knockdown in Brx-82 HER2⁺ CTCs increases the expression of NRF2-driven cytoprotective genes downstream of the Notch pathway. Representative data of two independent experiments are shown; s.e.m. (error bars). **e**, Quantitation of the interconversion of HER2⁺ cells from single-cell clones into 5- to 9-cell and >10-cell clusters following treatment with 10mM H₂O₂; *t*-test, $P < 0.05$; $n = 10$. **f**, Paclitaxel treatment of mice with tumours derived

from Brx-142 FACS-purified HER2⁺ CTCs, demonstrating a reduction in CTCs, and HER2[−] CTCs with no change in counts; *t*-test $P < 0.05$; NS, not significant. **g**, Paclitaxel treatment of mice with mammary xenografts derived from parental CTC line Brx-142 showing initial tumour response, followed by recurrent tumour growth. IHC analysis and quantitation of the recurrent tumour shows greatly reduced HER2⁺ (brown stain) cell composition in the Paclitaxel drug treated (T, 3 weeks post-treatment) tumour compared with the untreated tumour U, and the recovered tumour (R, 5 weeks post-treatment). Bar indicates duration of drug treatment (Rx). Scale bar, 100 μ m; two-way ANOVA, $P < 0.0001$; $n = 6$. Representative images from five independent fields per tumour are shown and quantified; *t*-test, $P < 0.001$. **h**, Dual GFP (red, cytoplasmic stain) and HER2 (brown, cell surface stain) IHC of tumour xenografts derived from mixed GFP-tagged HER2⁺ and untagged HER2[−] CTC cultures demonstrating enhanced conversion from GFP⁺/HER2⁺ to GFP⁺/HER2[−] after 4 weeks of paclitaxel treatment; *t*-test, $P < 0.0001$; $n = 6$. Scale bar, 100 μ m. Arrows indicate interconverting cells. Representative images from five independent fields per tumour are shown. **i**, Mouse tumour xenografts derived from the CTC line Brx-142 treated with a combination of the Notchi³ (LY-414575) and paclitaxel shows diminished tumour relapse; $n = 6$. Bar indicates treatment duration.

An endosomal tether undergoes an entropic collapse to bring vesicles together

David H. Murray^{1*}, Marcus Jahnel^{1,2,3*}, Janelle Lauer¹, Mario J. Avellaneda^{1,2†}, Nicolas Brouilly¹, Alice Cezanne¹, Hernán Morales-Navarrete¹, Enrico D. Perini^{1,2}, Charles Ferguson⁴, Andrei N. Lupas⁵, Yannis Kalaidzidis¹, Robert G. Parton^{4,6}, Stephan W. Grill^{1,2,3} & Marino Zerial¹

An early step in intracellular transport is the selective recognition of a vesicle by its appropriate target membrane, a process regulated by Rab GTPases via the recruitment of tethering effectors^{1–4}. Membrane tethering confers higher selectivity and efficiency to membrane fusion than the pairing of SNAREs (soluble N-ethylmaleimide-sensitive factor attachment protein receptors) alone^{5–7}. Here we address the mechanism whereby a tethered vesicle comes closer towards its target membrane for fusion by reconstituting an endosomal asymmetric tethering machinery consisting of the dimeric coiled-coil protein EEA1 (refs 6, 7) recruited to phosphatidylinositol 3-phosphate membranes and binding vesicles harbouring Rab5. Surprisingly, structural analysis reveals that Rab5:GTP induces an allosteric conformational change in EEA1, from extended to flexible and collapsed. Through dynamic analysis by optical tweezers, we confirm that EEA1 captures a vesicle at a distance corresponding to its extended conformation, and directly measure its flexibility and the forces induced during the tethering reaction. Expression of engineered EEA1 variants defective in the conformational change induce prominent clusters of tethered vesicles *in vivo*. Our results suggest a new mechanism in which Rab5 induces a change in flexibility of EEA1, generating an entropic collapse force that pulls the captured vesicle towards the target membrane to initiate docking and fusion.

EEA1, as nearly all putative coiled-coil tethering proteins, extends more than ten times the length of SNARE proteins^{8,9}. To explain how such a long molecule can mediate membrane tethering but also allow the membranes to come closer for fusion, we reconstituted a minimal asymmetric membrane tethering in liposomes containing EEA1, Rab5 and different fluorescent tracers (Fig. 1a and Extended Data Fig. 1b–e). EEA1 binds to phosphatidylinositol 3-phosphate (PI(3)P) via its carboxy (C) terminus with high affinity (dissociation constant $K_d \approx 50$ nM)^{7,10–12}, and to Rab5:GTP via its amino (N) terminus with comparatively lower affinity ($K_d \approx 2.4$ μ M)¹³. Liposomes containing PI(3)P and labelled with RhoDPPE effectively recruited EEA1 and tethered to DiD-labelled Rab5-6 \times His-liposomes, as analysed by confocal microscopy (Fig. 1a–c). The reaction required EEA1, Rab5 and GTP- γ S, as no co-localization was observed in the presence of GDP. The efficiency of tethering approached that of biotin–streptavidin liposomes (Fig. 1d). Furthermore, no co-localization was observed between pairs of liposomes harbouring Rab5 (Fig. 1e). Therefore, Rab5, EEA1 and PI(3)P form a minimal endosomal asymmetric membrane tethering machinery.

In principle, the N terminus of EEA1 could also bind Rab5 *in cis*: that is, on the same membrane. However, the presence of Rab5 on both pairs of liposomes, as in early endosomes *in vivo*, did not interfere with the tethering activity of EEA1 *in vitro*, as tethering was

indistinguishable between the asymmetric and symmetric conditions (Fig. 1c, e). Moreover, coiled-coil prediction algorithms estimate a central segment of nearly ~ 200 nm (refs 14, 15) (Extended Data Fig. 1a), suggesting that the molecule adopts an extended conformation. Indeed, filamentous EEA1-positive structures emanating from the surface of early endosomes *in vivo* have been observed by electron microscopy¹¹. In further support of this interpretation, we visualized the N and C termini of EEA1 using specific antibodies by super-resolution microscopy in HeLa cells (Fig. 1f, g, Extended Data Fig. 1f–h and Methods). If the N terminus of EEA1 bound Rab5 *in cis*, it should co-localize with the C terminus. Strikingly, the ends of EEA1 could instead be resolved, with the N terminus extending radially from the C terminus into the cytoplasm. We estimated an end-to-end distance of 141 ± 47 nm (mean \pm s.d.; Fig. 1h), in the range of the predicted length and rigidity of coiled-coils.

To characterize the distances and dynamics of the tethering reaction, we generated bead-supported membranes (10 μ m silica microspheres) harbouring green fluorescent protein (GFP)–Rab5 (Fig. 1i and Extended Data Fig. 2). These tethered to liposomes containing PI(3)P in the presence of GTP- γ S but not GDP in an EEA1 concentration-dependent manner (Extended Data Fig. 2g, h). Time-lapse microscopy showed that some liposomes were captured by the bead-supported membrane, while others diffused away (Extended Data Fig. 2i and Supplementary Videos 1 and 2), similar to the behaviour of endosomes *in vivo*¹⁶. We next measured the distances between the tethered vesicle and GFP–Rab5 (Fig. 1j, Extended Data Fig. 2j and Methods). Surprisingly, we observed distances ranging from 20 nm up to approximately the predicted length of 200 nm (mean \pm s.d.; 84 ± 56 nm) (Fig. 1k). Such a broad distribution is irreconcilable with the predicted length of EEA1 and suggests that EEA1 may change its conformation.

We determined the conformation of EEA1 using rotary shadowing electron microscopy and image analysis (Fig. 2a). The measurements of contour length and mean end-to-end distance followed Gaussian distributions with an average of 222 ± 26 nm (Fig. 2b, top) and 195 ± 26 nm (Fig. 2b, bottom), respectively, confirming that the molecule is largely extended, as *in vivo*¹¹ (Fig. 1g, h). However, this is incompatible with the much shorter distances between tethered vesicles *in vitro* (Fig. 1k). Therefore, we asked whether binding to Rab5 may cause EEA1 to adopt a more compact conformation. Remarkably, this was the case. Addition of Rab5:GTP- γ S (Fig. 2c) resulted in a significant fraction of bent EEA1 molecules having a substantially reduced end-to-end distance of 122 ± 50 nm (Fig. 2d).

To gain further insights into this mechanism, we generated two mutants with alterations in the coiled-coil but retaining the Rab5- and PI(3)P-binding domains (Extended Data Fig. 3 and Methods). In the extended EEA1 mutant, we removed regions of discontinuity

¹Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstraße 108, 01307 Dresden, Germany. ²Biotechnology Center, Technical University Dresden, Tatzberg 47/49, 01307 Dresden, Germany. ³Max Planck Institute for the Physics of Complex Systems, Nöthnitzerstraße 38, 01187 Dresden, Germany. ⁴Institute for Molecular Bioscience, The University of Queensland, St Lucia 4072, Australia. ⁵Department of Protein Evolution, Max-Planck-Institute for Developmental Biology, 72076 Tübingen, Germany. ⁶Centre for Microscopy and Microanalysis, The University of Queensland, St Lucia 4072, Australia. [†]Present address: FOM Institute AMOLF, Science Park 104, 1098 XG Amsterdam, the Netherlands.

*These authors contributed equally to this work.

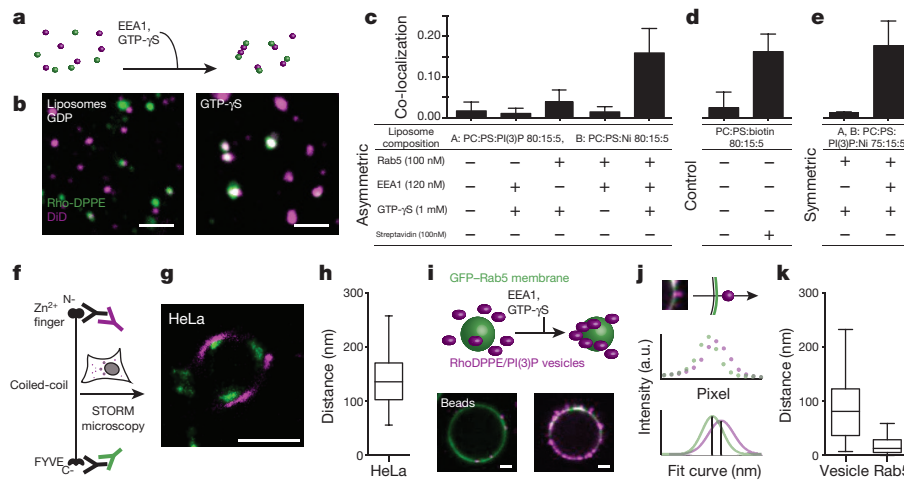


Figure 1 | EEA1, Rab5 and PI(3)P form an asymmetric tethering machinery. **a, b**, Vesicle–vesicle tethering assay. Rho-DPPE liposomes harbouring Rab5 (green) tether to DiD-PI(3)P liposomes (magenta) upon addition of EEA1 and GTP- γ S but not GDP (**a**, scheme; **b**, microscopy; representative of $n = 20$). Scale bar, 2 μ m. **c–e**, Analysis of vesicle co-localization. Asymmetric (**c**) and symmetric (**e**) tethering required Rab5, PI(3)P and EEA1, streptavidin–biotin control (**d**) (mean \pm s.d., $n = 3$). **f–h**, *In vivo* stochastic optical reconstruction microscopy (STORM) defines the extension of EEA1. The N-terminal (magenta) and C-terminal (green) domains of EEA1 (**f**) were differentially labelled. Representative

STORM image (**g**, of $n = 22$) and quantification of EEA1 extension (**h**, box–whisker plot with median, 25/75 quartiles and minimum/maximum error bars, $n = 86$, representative experiment) from endosomes. Scale bar, 500 nm. **i**, Bead-supported membrane tethering similar to **a** and **b**. Representative of $n = 20$. Scale bar, 2 μ m. **j, k**, Distance of tethered vesicles (magenta) from the membrane (green). The intensity per pixel was plotted, fitted to determine the relative distances and quantified (**k**) (vesicle-membrane and Rab5-membrane, representative experiment; box–whisker plot as in **h**, mean \pm s.d., $n = 36$ and 14).

between heptad repeats creating a more idealized, extended coiled-coil. In the swapped EEA1 mutant, we swapped the coiled-coil regions between the N and C termini. Electron microscopy analysis revealed that the extended mutant was impaired in the Rab5-induced conformational change (Fig. 2i and Extended Data Fig. 4a–c). In contrast, the swapped mutant was mostly bent, often presented kinks, and did not significantly change conformation upon Rab5 binding (Fig. 2f and Extended Data Fig. 4e–g). These results suggest that coiled-coil discontinuities and their physical arrangement are

critical for the structure of EEA1 and its Rab5-induced conformational change.

To shed light on how EEA1 adopts a compact conformation upon Rab5 binding, we measured the curvature along the contour of molecules. We aligned N-terminally MBP-tagged EEA1 and determined how the tangents to the contour change by 8 nm steps along the contour (Methods and Extended Data Fig. 5). Interestingly, the variance of this measure of curvature calculated over the ensemble of molecules increased significantly upon Rab5:GTP- γ S binding (Fig. 2g), indicating

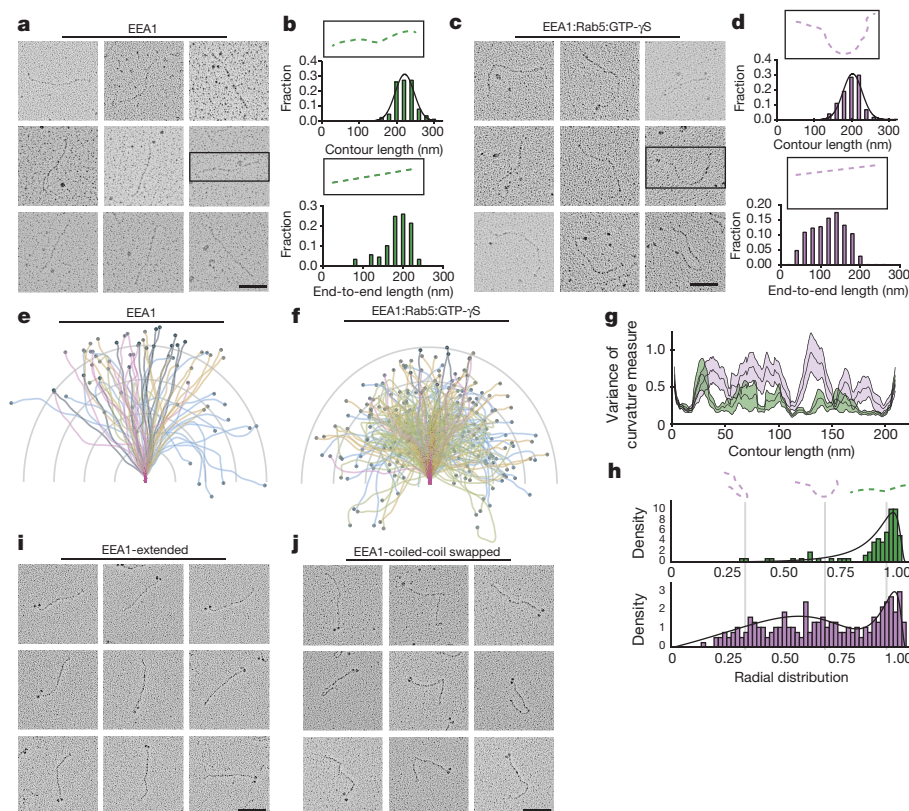


Figure 2 | EEA1 changes flexibility upon Rab5 binding. **a, c, i, j**, Representative examples of rotary-shadowing electron microscopy of EEA1 (**a**), EEA1 + Rab5:GTP- γ S (**c**), EEA1-extended (**i**) and -swapped (**j**) variants. Scale bar, 100 nm; $n = 88$, $n = 212$, $n = 90$, $n = 145$, respectively. **b, d**, Contour and end-to-end length histograms for EEA1 (green, $n = 88$) and EEA1 + Rab5:GTP- γ S (magenta, $n = 212$). **e, f**, Visual comparison of aligned EEA1 proteins. The highlighted ends of EEA1 + Rab5:GTP- γ S lie significantly closer to the origin. Hemispheres demarcate 50 nm. **g**, Variance of curvature measures along the contour of aligned EEA1 + Rab5:GDP (green) and EEA1 + Rab5:GTP- γ S (magenta) molecules ($n = 90$, $n = 145$, respectively). **h**, Radial distribution functions define the extension probability for EEA1 \pm Rab5:GTP- γ S (–Rab5:GTP- γ S, green; +Rab5:GTP- γ S, magenta) with fit (black lines).

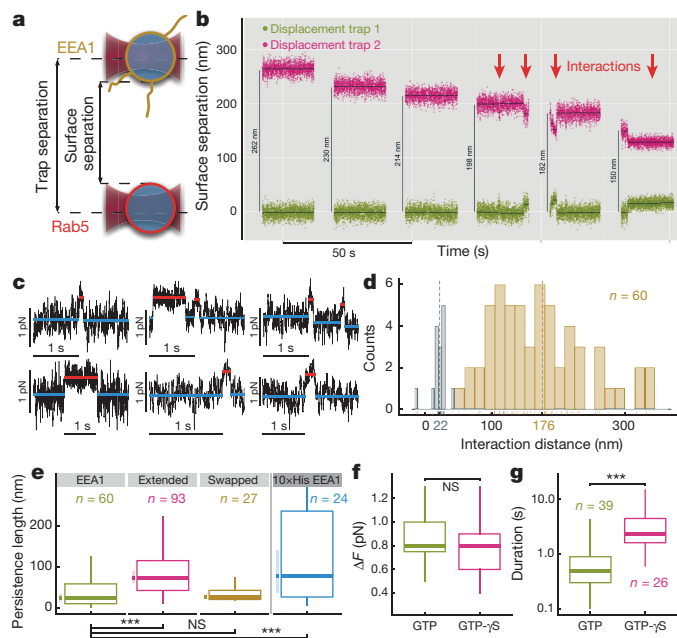


Figure 3 | EEA1 collapse generates a force. **a**, Scheme of bead-supported membranes harbouring EEA1 or Rab5 captured by dual-trap optical tweezers. **b**, **c**, Traps moved successively closer until interactions (arrows) were observed, characterized by increase in force and decrease in variance (c). **d**, Interaction distance consistent with length of extended EEA1. Silica microspheres (negative control) in grey. **e**, Persistence length distributions of EEA1 and variants from optical tweezers measurements. **f**, Force did not depend on GTP hydrolysis ($P > 0.15$); $n = 39, 26$ respectively. **g**, Interaction duration (log-scale) was prolonged by GTP- γ S ($P < 10^{-4}$). Mann–Whitney–Wilcoxon test (e–g); box–whisker plot with Tukey error bars (e–g).

that EEA1 displays a larger variety of curvatures upon Rab5:GTP binding. Such changes occurred along the entire length of the molecule, with some regions increasing in flexibility more than others (Fig. 2g), but were not observed for the EEA1 mutants (Extended Data Fig. 5f–i).

Although molecules are adsorbed onto a 2D surface, some aspects of their 3D conformations are captured (Methods). Analysis of the kurtosis of the distribution of angles between contour tangents indicated that 3D shape fluctuations are retained for the entire contour of EEA1 in the presence of Rab5:GDP, but only up to 60 nm with Rab5:GTP- γ S (Methods and Extended Data Fig. 6). Moreover, tangent–tangent correlations of the contour in this regime revealed that Rab5:GTP- γ S binding results in a faster decay. Generally, the worm-like chain (WLC) model is used to describe fluctuations in polymer shapes and

capture aspects of the physics underlying their shape fluctuations¹⁷ (Methods). In the WLC model, the polymer is considered a homogeneous molecule with its flexibility determined by a bending stiffness reflected in a characteristic length, the persistence length, over which correlations between tangents to the contour decay. We applied the WLC model to EEA1 and determined an effective persistence length of 246 ± 42 nm for the unbound and 74 ± 3 nm for the Rab5:GTP- γ S-bound ensembles. In contrast, the extended EEA1 mutant had similar effective persistence lengths in either state (unbound = 183 ± 13 nm and bound = 224 ± 25 nm; Supplementary Data Table).

To corroborate these estimates, we fitted the radial distribution functions (that is, the probability of observing a given end-to-end distance) of the molecules extracted from the electron microscopy data with analytical solutions of the WLC model¹⁸ (Methods). This showed a clear reduction in effective persistence length of EEA1 upon Rab5:GTP binding (Fig. 2h). In contrast, the extended EEA1 mutant maintained a similar radial distribution regardless of Rab5 (Extended Data Fig. 4d).

Reducing the persistence length of EEA1 makes the molecule flexible. However, the tether is still extended and, therefore, in an out-of-equilibrium conformation (Fig. 2e). As a result, it will undergo an entropic collapse, with its end-to-end distance decreasing towards a new equilibrium (Fig. 2f). This process generates a force that could pull the membranes together (estimated ~ 3 pN (Methods)). In some sense, the extended molecule is like a loaded spring that rapidly recoils upon Rab5 binding.

To provide experimental evidence for entropic collapse of EEA1, we made use of high-resolution dual-trap optical tweezers (Methods). Two glass $2\mu\text{m}$ microspheres coated with membranes were held in optical traps (Fig. 3a). One trap was moved closer to the other, in iterative cycles of approaching, pausing and retracting (Fig. 3b). At distances below 250 nm and at low concentrations of EEA1 (5–40 nM) to ensure single-molecule events, we observed transient interactions as a decrease in the mean and variance of the distance between the two beads (Fig. 3b, red arrows, Fig. 3c and Extended Data Fig. 7a, d). Interactions were infrequent, as expected for single molecules and non-existent without EEA1, whereas their frequency and duration increased at high concentrations of EEA1 (400 nM) (Extended Data Fig. 7e and Methods). The interaction distance was broad (Fig. 3d), with the mean 176 ± 76 nm comparing favourably with rigid EEA1 (Fig. 2b).

To test the prediction that EEA1 becomes flexible upon Rab5 binding, for each tethered molecule we determined its effective persistence length from the capture distance, and measured force increase (Fig. 3c) and bead displacements using the WLC model (Methods). Strikingly, we obtained a median effective persistence length of 23 ± 10 nm (Fig. 3e). For more than 80% of the molecules the persistence length was no more than half of the contour length, confirming that Rab5-bound

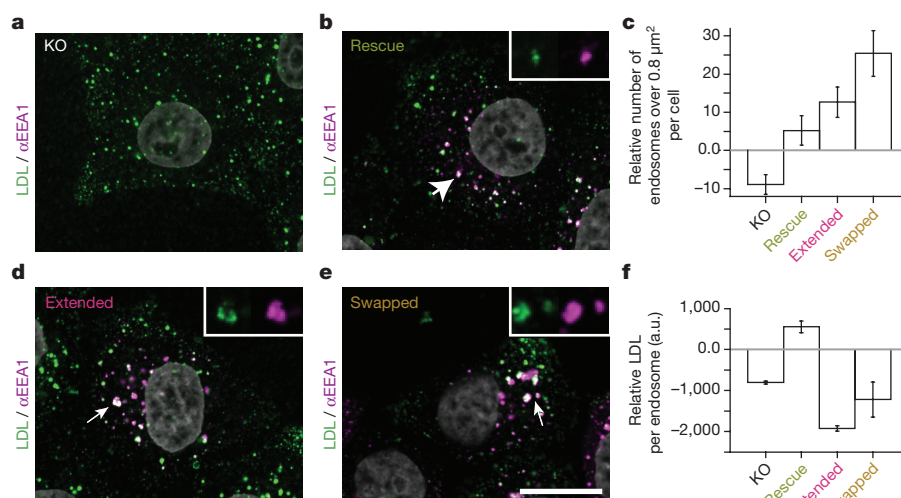


Figure 4 | EEA1 mutants blocking entropic collapse induce trafficking defects. **a**, **b**, **d**, **e**, Confocal images of HeLa EEA1-KO cells (a), rescued with EEA1, extended or swapped mutants (b, d, e). Uptake of LDL (green) and immunostaining for EEA1 (magenta). Inset, endosomes depicted at arrows. Representative of $n = 30$ images per condition (Methods). Scale bar, $10\mu\text{m}$. **c**, **f**, Relative difference in number of large endosomes (c) and LDL fluorescence (f) (a.u., arbitrary units). Mean \pm s.d., representative experiment of 3, $n = 30$ images. $P < 0.01$ versus HeLa, t -test, except rescue.

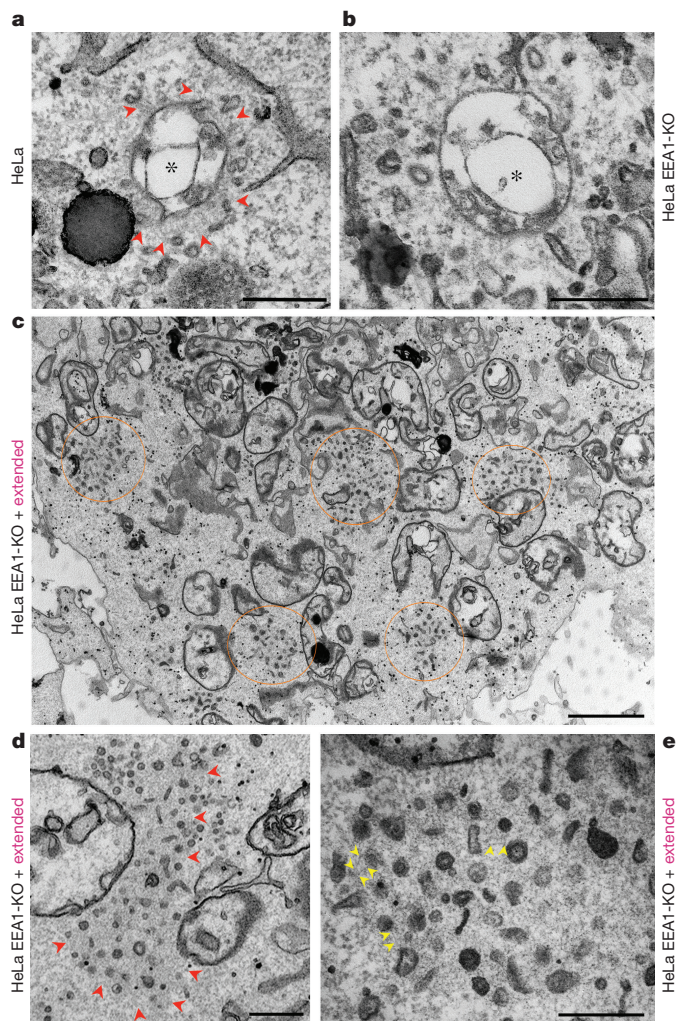


Figure 5 | Ultrastructural analysis of EEA1 KO and mutant rescue cells. **a**, Dense filamentous network (arrowheads) around an early endosome (asterisks) in HeLa. Many smaller vesicular or tubular profiles were consistently observed at the network periphery. Representative of $n = 33$. **b**, A filamentous network was less prominent in HeLa EEA1-KO with no obvious concentration of vesicles near the endosomal surface. Representative of $n = 54$. **c–e**, HeLa EEA1-KO expressing the extended EEA1 variant showed clusters of vesicles throughout the cytoplasm and no classical endosomal morphology. The clusters were clearly delineated by a zone of cytoplasm with distinct density (circled areas). Higher magnification revealed fine wispy material surrounding the clustered vesicles (**d**, **e**; arrowheads) and evidence of discrete filaments (between the arrowheads in **e**). Representative of $n = 56$. Scale bars: **a**, **b**, **d**, **e**, 500 nm; **c**, 2 μm .

EEA1 is flexible. In contrast, the extended EEA1 mutant remained significantly more rigid than EEA1 (Fig. 3e). Rab5 binding is necessary to trigger structural and conformational changes on EEA1. When Rab5 was bypassed by His-tag-mediated tethering, EEA1 flexibility was significantly lower than that of EEA1 with Rab5 (Fig. 3e).

If EEA1 becomes flexible upon capture, an entropic pulling force will be generated. This entropic force balances with the force exerted by the optical traps as the molecule undergoes the collapse and as the system finds its new equilibrium (Extended Data Fig. 7h)¹⁹. For a capture distance of 195 nm and a peak collapse force of 3 pN, we predict a force balance at ~ 0.6 pN (Methods), consistent with our tweezer measurements of 0.5 ± 0.3 pN (Fig. 3c). EEA1 binding to Rab5 requires the GTP-bound form. No significant force differences were observed in the presence of the non-hydrolysable analogue GTP- γ S or GTP (Fig. 3f). In contrast, the duration of the interaction was much

prolonged (Fig. 3g), as expected given that GTP- γ S stabilizes Rab5 in the active form²⁰. Finally, replacing EEA1-Rab5 binding with $10\times$ His-EEA1 tethering to Ni-NTA-beads resulted in a decreased collapse force (Extended Data Fig. 7i).

To validate *in vivo* the mechanism observed *in vitro*, we genome-edited HeLa cells to disrupt the EEA1 gene (HeLa EEA1-KO; Fig. 4a, Extended Data Fig. 8c and Methods), and analysed the distribution of Rab5-positive endosomes and the uptake of cargo (low-density lipoprotein (LDL)) by confocal microscopy (Fig. 4a). HeLa EEA1-KO displayed a significant reduction in Rab5 endosome size, particularly for the largest endosomes (Fig. 4c), and a marked decrease in cargo (LDL) uptake (Fig. 4f). Expression of EEA1 rescued the normal, rounded morphology of endosomes (Fig. 4b and Extended Data Fig. 8f, i) and LDL uptake (Fig. 4c). In contrast, the expression of both extended and swapped EEA1 mutants generated enlarged endosomes and inhibited cargo uptake (Fig. 4c–f).

Because the size of endosomes is below the resolution limit of light microscopy, we performed electron microscopy on the HeLa EEA1-KO cells (Fig. 5 and Extended Data Fig. 9). The filamentous material on endosomes¹¹ was much reduced in HeLa EEA1-KO cells (Fig. 5a, b, and Extended Data Fig. 8n) and restored by the re-expression of EEA1 on endosomes that appeared normal or enlarged, consistent with the light microscopy analysis (Fig. 4b). Strikingly, cells expressing the extended EEA1 mutant had large ($>1\mu\text{m}$) clusters of small vesicles, within areas filled with filamentous material (Fig. 5d, e), suggesting that they are arrested in a tethered state (Fig. 4d, e). The distance between the tethered vesicles was significantly longer than that between endosomes in control cells (Extended Data Fig. 8o), consistent with the mutant EEA1 being incapable of undergoing entropic collapse to shorter distances (Figs 2e and 3e). Similar endosomal clusters were induced by the swapped mutant (Extended Data Fig. 8m).

Our data suggest a new mechanochemical cycle of EEA1 regulated by Rab5:GTP binding and GTP hydrolysis. On early endosomes, EEA1 is in the extended state (Fig. 2e) and increases the probability of capturing a vesicle bearing Rab5. Similarly, it forms a Rab5-selectivity barrier (analogous to a polymer brush)²¹. When Rab5 on an incoming vesicle binds EEA1, it induces an allosteric conformational change, from extended to flexible (Fig. 2f). This shows a new function of Rab proteins beyond effector recruitment. The reduction in persistence length of EEA1 causes its entropic collapse, releasing up to $\sim 14 k_B T$ of mechanical energy (Extended Data Fig. 7k) and generating up to 3 pN of force that could pull the vesicle closer to its target membrane where it may diffuse²² or be brought by other Rab5 effectors^{23,24} within the range of trans-SNARE pairing. This mechanism explains why the Rab5 machinery dramatically increases the efficiency of SNARE-mediated membrane fusion²³. The mechanical energy released by EEA1 is of the order of the free energy released by GTP hydrolysis. However, the energy required to complete the cycle could potentially also come from chaperones.

A key question is how Rab5 can induce such a long-range allosteric effect. This is not uncommon among coiled-coil proteins^{25,26}. The entropic collapse mechanism is different, however, for other membrane tethering factors²⁷. In the course of this study, the GCC185 tether was shown to bend through central joints²⁷. For EEA1, instead (1) the arrangement and structure of the coiled-coils and (2) Rab5 binding are critical for the propagation of allosteric conformational changes (Extended Data Fig. 10). We can envisage different mechanisms (see Supplementary Discussion), such as local register shifts. In dynein, dynamics in the heptad register prove critical to functionally link ATP binding and microtubule binding at opposite ends of its coiled-coil stalk^{28,29}. Further ad hoc structural studies are necessary to resolve this outstanding problem. The entropic collapse upon stiffness reduction could be an effective and general mechanism used not only by membrane tethers but also by many coiled-coil proteins for generating an attractive force in diverse biological processes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 September 2015; accepted 19 July 2016.

Published online 24 August 2016.

1. Bröcker, C., Engelbrecht-Vandré, S. & Ungermann, C. Multisubunit tethering complexes and their role in membrane fusion. *Curr. Biol.* **20**, R943–R952 (2010).
2. Brown, F. C. & Pfeffer, S. R. An update on transport vesicle tethering. *Mol. Membr. Biol.* **27**, 457–461 (2010).
3. Zerial, M. & McBride, H. Rab proteins as membrane organizers. *Nature Rev. Mol. Cell Biol.* **2**, 107–117 (2001).
4. Munro, S. Organelle identity and the organization of membrane traffic. *Nature Cell Biol.* **6**, 469–472 (2004).
5. Mayer, A. & Wickner, W. Docking of yeast vacuoles is catalyzed by the Ras-like GTPase Ypt7p after symmetric priming by Sec18p (NSF). *J. Cell Biol.* **136**, 307–317 (1997).
6. Christoforidis, S., McBride, H. M., Burgoyne, R. D. & Zerial, M. The Rab5 effector EEA1 is a core component of endosome docking. *Nature* **397**, 621–625 (1999).
7. Rubino, M., Miaczynska, M., Lippé, R. & Zerial, M. Selective membrane recruitment of EEA1 suggests a role in directional transport of clathrin-coated vesicles to early endosomes. *J. Biol. Chem.* **275**, 3745–3748 (2000).
8. Gao, Y. *et al.* Single reconstituted neuronal SNARE complexes zipper in three distinct stages. *Science* **337**, 1340–1343 (2012).
9. Kiessling, V. & Tamm, L. K. Measuring distances in supported bilayers by fluorescence interference-contrast microscopy: polymer supports and SNARE proteins. *Biophys. J.* **84**, 408–418 (2003).
10. Dumas, J. J. *et al.* Multivalent endosome targeting by homodimeric EEA1. *Mol. Cell* **8**, 947–958 (2001).
11. Wilson, J. M. *et al.* EEA1, a tethering protein of the early sorting endosome, shows a polarized distribution in hippocampal neurons, epithelial cells, and fibroblasts. *Mol. Biol. Cell* **11**, 2657–2671 (2000).
12. Simonsen, A. *et al.* EEA1 links PI(3)K function to Rab5 regulation of endosome fusion. *Nature* **394**, 494–498 (1998).
13. Mishra, A., Eathiraj, S., Corvera, S. & Lambright, D. G. Structural basis for Rab GTPase recognition and endosome tethering by the C2H2 zinc finger of early endosomal autoantigen 1 (EEA1). *Proc. Natl Acad. Sci. USA* **107**, 10866–10871 (2010).
14. Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164 (1991).
15. McDonnell, A. V., Jiang, T., Keating, A. E. & Berger, B. Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics* **22**, 356–358 (2006).
16. Rink, J., Ghigo, E., Kalaidzidis, Y. & Zerial, M. Rab conversion as a mechanism of progression from early to late endosomes. *Cell* **122**, 735–749 (2005).
17. Landau, L. D. & Lifshitz, E. M. *Statistical Physics* 3rd edn, Part 1, Vol. 5, Ch. 12, 396–400 (Butterworth-Heinemann, 1980).
18. Wilhelm, J. & Frey, E. Radial distribution function of semiflexible polymers. *Phys. Rev. Lett.* **77**, 2581–2584 (1996).
19. Otto, O., Sturm, S., Laohakunakorn, N., Keyser, U. F. & Kroy, K. Rapid internal contraction boosts DNA friction. *Nature Commun.* **4**, 1780 (2013).
20. Rybin, V. *et al.* GTPase activity of Rab5 acts as a timer for endocytic membrane fusion. *Nature* **383**, 266–269 (1996).
21. Milner, S. T. Polymer brushes. *Science* **251**, 905–914 (1991).
22. Degtyar, V. E., Allersma, M. W., Axelrod, D. & Holz, R. W. Increased motion and travel, rather than stable docking, characterize the last moments before secretory granule fusion. *Proc. Natl Acad. Sci. USA* **104**, 15929–15934 (2007).
23. Ohya, T. *et al.* Reconstitution of Rab- and SNARE-dependent membrane fusion by synthetic endosomes. *Nature* **459**, 1091–1097 (2009).
24. Perini, E. D., Schaefer, R., Stöter, M., Kalaidzidis, Y. & Zerial, M. Mammalian CORVET is required for fusion and conversion of distinct early endosome subpopulations. *Traffic* **15**, 1366–1389 (2014).
25. Moreno-Herrero, F. *et al.* Mesoscale conformational changes in the DNA-repair complex Rad50/Mre11/Nbs1 upon binding DNA. *Nature* **437**, 440–443 (2005).
26. Taylor, K. C. *et al.* Skip residues modulate the structural properties of the myosin rod and guide thick filament assembly. *Proc. Natl Acad. Sci. USA* **112**, E3806–E3815 (2015).
27. Cheung, P. Y., Limouse, C., Mabuchi, H. & Pfeffer, S. R. Protein flexibility is required for vesicle tethering at the Golgi. *eLife* **4**, e12790 (2015).
28. Schmidt, H., Zalyte, R., Urnavicius, L. & Carter, A. P. Structure of human cytoplasmic dynein-2 primed for its power stroke. *Nature* **518**, 435–438 (2015).
29. Kon, T. *et al.* Helix sliding in the stalk coiled coil of dynein couples ATPase and microtubule binding. *Nature Struct. Mol. Biol.* **16**, 325–333 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank R. Schäfer for project support. We acknowledge discussions with J. Morin, Ü. Coskun, A. Honigsmann, S. Sturm and T. Leonard, and F. Jülicher, E. Schäfer and K. Simons for reading the manuscript. We thank M. Brandstetter and the electron microscopy facility of the Vienna Biocenter. We thank the Light Microscopy, Protein Expression, Chromatography, and High-throughput Technology Development Studio of the Max Planck Institute of Molecular Cell Biology and Genetics. During part of the work, M.J. was supported by a PhD scholarship of the Böhlinger Ingelheim Fonds. M.J.A. was supported by the La Caixa and Deutscher Akademischer Austauschdienst scholarship. R.P. was supported by the National Health and Medical Research Council of Australia (program grant APP1037320 and Senior Principal Research Fellowship 569452) and the Australian Research Council Centre of Excellence (CE140100036). We acknowledge the Australian Microscopy & Microanalysis Research Facility at the Center for Microscopy and Microanalysis at The University of Queensland. S.W.G. was supported by the Deutsche Forschungsgemeinschaft (SPP 1782, GSC 97, GR 3271/2, GR 3271/3, GR 3271/4), the European Research Council (grant number 281903) and the Human Frontier Science Program (RGP0023/2014). This research was supported by the Max Planck Society and funds of the Deutsche Forschungsgemeinschaft (Transregio 83).

Author Contributions D.H.M., M.J., S.W.G. and M.Z. conceived the project together. D.H.M. prepared all reagents, performed experiments and their analysis. M.J. and S.W.G. interpreted data in the context of polymer physics. M.J. performed optical tweezer experiments with D.H.M. M.J.A. and E.P. performed initial tweezer experiments. M.J. and M.J.A. analysed tweezer experiments. D.H.M., J.L. and A.N.L. designed mutants. N.B. performed super-resolution experiments. A.C. assisted in reconstitution experiments. C.F. and R.G.P. performed cell electron microscopy, and D.H.M., H.M.-N. and M.J. analysed electron microscopy data. Y.K. analysed cell microscopy. D.H.M., M.J., S.W.G. and M.Z. wrote the manuscript with input from all the authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.W.G. (stephan.grill@biotec.tu-dresden.de) or M.Z. (Zerial@mpi-cbg.de).

Reviewer Information *Nature* thanks C. Schmidt, J. Zimmerberg and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Statistics. Sample size was not predetermined. For cell electron microscopy, samples were double-blind examined. Other experiments were not randomized or blinded. Box-whisker plots all show median, 25/75 quartiles by box boundaries and minimum/maximum values by errors, with the exception of Fig. 3 and Extended Data Fig. 7 which use Tukey-defined error bars.

Cloning, expression and purification of proteins. Human Rab5-6 \times His and GFP-Rab5-6 \times His were expressed and purified essentially as previously described in the *Escherichia coli* expression system⁶. Human Rabex-5 amino-acid residues 131–394 were PCR and restriction cloned into a pGST-parallel2 vector containing a TEV cleavable N-terminal glutathione-S-transferase (GST)^{29,30}. Expression and purification was performed essentially as described³¹. Briefly, *E. coli*-expressed proteins were transformed into BL21(DE3) cells and grown at 37 °C until absorbance at 600 nm ($A_{600\text{ nm}}$) of 0.8, whereupon the incubator was reduced to 18 °C. After 30 min, cultures were induced with 0.1 mM IPTG and grown overnight (16 h). Cell pellets were resuspended in standard buffer (20 mM Tris pH7.4, 150 mM NaCl, 0.5 mM TCEP) and flash frozen in liquid nitrogen. All subsequent steps performed at 4 °C or on ice. Cell pellets were resuspended in standard buffer supplemented with 1 mM MgCl₂ for GTPases, and protease inhibitor cocktail (chymostatin 6 μ g/ml, leupeptin 0.5 μ g/ml, antipain-HCl 10 μ g/ml, aprotinin 2 μ g/ml, pepstatin 0.7 μ g/ml, APMSF 10 μ g/ml), homogenized and lysed by sonication. Histidine-tagged proteins were bound in batch to Ni-NTA resin in the presence of 20 mM imidazole, and eluted with 200 mM imidazole. GST-tagged proteins were purified on GS resin (GS-4B, GE Healthcare) by binding for 2 h followed by stringent washing, and cleavage from resin overnight. Imidazole-containing samples were immediately diluted after elution and tags cleaved during overnight dialysis. Following dialysis and tag cleavage, samples were concentrated and TEV or HRV 3C protease was removed by reverse purification through Ni-NTA or GS resin. Samples were then purified by size-exclusion chromatography on Superdex 200 columns in standard buffer.

Human EEA1 was purified as a GST fusion in a pOEM series vector (Oxford Expression Technologies) modified to contain a HRV 3C-cleavable N-terminal GST and protease cleavage site or from a modified pFastbac1 vector (Thermo Fisher Scientific)²³. Some samples were also purified as 6 \times His-MBP and 10 \times His fusions from a modified pOEM vector (rotary shadowing for N-to-C terminus alignment, and optical tweezer control, respectively; all other experiments performed with tags removed). Mutants were purified identically to wild-type EEA1.

SF9 cells growing in ESF921 media (Expression Systems) were co-transfected with linearized viral genome and the expression plasmid and selected for high infectivity. P1 and P2 virus was generated according to the manufacturer's protocol, and expression screens and time courses performed to optimize expression yield. Best viruses were used to infect 1–2 l SF9 cells at 10⁶ cells/ml at 1% vol/vol and routinely harvested after 40–48 h at about 1.5 \times 10⁶ cells/ml, suspended in standard buffer and flash frozen in liquid nitrogen. Pellets were thawed on ice and lysed by Dounce homogenizer. Purification took place rapidly in standard buffer at 4 °C on GS resin in batch format. Bound protein was washed thoroughly and cleaved from resin by HRV 3C protease overnight. Proteins retaining 6 \times His-MBP tags were purified on amylose resin and eluted with 10 mM maltose. Protein retaining 10 \times His were eluted from Ni-NTA resin in standard buffer supplemented with 200 mM imidazole. All EEA1 and mutants were immediately further purified by Superose 6 size-exclusion chromatography where they eluted as a single peak. All experiments were performed with a preparation confirmed for Rab5 and PI(3)P binding. Concentrations were determined by UV280 and Bradford assay. All proteins were aliquoted and flash frozen in liquid nitrogen and stored at –80 °C.

EEA1 variants extended and swapped were synthesized genes optimized for insect cell expression (Genscript). The extended mutant has regions of low coiled-coil prediction removed, resulting in an EEA1 construct 1,286 amino acids in length (versus 1,411 in wild-type EEA1) (see Extended Data Fig. 3). The swapped mutant has the C-terminal portion of the coiled-coil rearranged to follow the N-terminal Zn²⁺-finger domains, and the N-terminal portion of the coiled-coil therefore rearranged to the C-terminal region of EEA1. Variants were treated identically to wild-type EEA1 in purification.

Static light scattering. An autosampler equipped Viskotek TDAMax system was used to analyse the light-scattering from purified EEA1. Sample was loaded the autosampler and passed through a TSKGel G5000PW column (Tosoh Biosciences) and fractions were subjected to scattering data acquisition. Data obtained were averaged across the protein elution volume and molecular masses determined in OmniSEC software package.

Lipids. The following lipids were purchased and used directly: DOPC, DOPS, DOGS-NiNTA, RhoDPPE (Avanti), DiD (Invitrogen) and PI(3)P (Echelon Biosciences). Lipids were dissolved in chloroform, except PI(3)P in 1:2:0.8 CHCl₃:MeOH:H₂O. All were stored at –80 °C.

Rab5/PI(3)P binding by EEA1. Early endosome fusion assay was performed as previously described³². To assess the ability of EEA1 to bind competently in a GTP-dependent manner to Rab5, Rab5 was bound to GS resin and subsequently loaded with nucleotide (GDP, GTP- γ S) as previously described⁶. Binding of EEA1 and all variants to immobilized Rab5 proceeded for 1 h at room temperature, and the washed Rab5 resin was evaluated for EEA1 binding by western blot. Similarly, the binding of EEA1 to PI(3)P containing liposomes was evaluated as previously described by formation of liposomes composed of DOPC:DOPS or DOPC:DOPS:PI(3)P (85:15 or 80:15:5 respectively)³³. Briefly, liposomes were formed from the hydration of lipids at 1 mM in standard buffer, and combined with EEA1 for 1 h before ultracentrifugation to separate supernatant and pellet for western blotting to evaluate EEA1 sedimentation. Rabbit anti-EEA1 antibody was made in our laboratory.

Preparation of liposomes. Liposomes were formed by extrusion as previously described³⁴. Liposome compositions for fluorescence microscopy tethering assays were DOPC:DOPS:DOGS-NiNTA, DOPC:DOPS:PI(3)P, DOPC:DOPS:biotin-DPPE, with RhoDPPE and DiD where applicable. Liposome compositions for bead-supported membranes were DOPC:DOPS:DOGS-NiNTA, DOPC:DOPS:PI(3)P. Solvent was evaporated under nitrogen and vacuum overnight. The resulting residue was suspended in standard buffer, rapidly vortexed, freeze-thawed five times by submersion in liquid N₂ followed by water at 40 °C, and extruded by 11 passes through two polycarbonate membranes with a pore diameter of 100 nm (Avestin). Vesicles stored at 4 °C were used within 5 days.

Bead-supported bilayer preparation. Silica beads (2 μ m NIST-traceable size-standards for optical tweezers, or 10 μ m standard microspheres for microscopy; Corpucular) were thoroughly cleaned in pure ethanol and Hellmanex (1% sol., Hellma Analytics) before storage in water. Supported bilayers were formed as previously described with modifications³⁵. Liposomes composed of DOPC:DOPS 85:15 (with 5% PI(3)P and DOGS-NiNTA where applicable) were added to a solution containing 250 mM NaCl for tethering assays (10 μ m) and 100 mM for optical tweezers (2 μ m), and 5 \times 10⁶ beads. Liposomes were added to final concentration of 100 μ M and incubated for 30 min (final volume 100 μ l). Samples were washed with 20 mM Tris pH7.4 three times by addition of 1 ml followed by gentle centrifugation (at 380g). Final wash was with standard buffer. Salt concentrations were optimized by examination of homogeneity at the transverse plane followed by examination of the excess membrane at the coverslip plane (see Extended Data Fig. 2a–d). We found that the membranes were extremely robust in conditions where the bilayer is fully formed, and could be readily pipetted and washed, consistent with previous reports³⁶. Membrane-coated beads were used within 1 h of production and always stored before use on a rotary suspension mixer.

Confocal microscopy of vesicle-vesicle tethering assay. Glass coverslips were cleaned in ethanol, Hellmanex and thoroughly rinsed in water. In these experiments, the following concentrations were used: 1 nM Rabex-5 (131–394), 100 nM Rab5-6 \times His, 120 nM EEA1. Experiments were performed in standard buffer with 5 mM MgCl₂ and 1 μ M nucleotide. Liposomes and proteins were pre-mixed in low-binding tubes at concentrations indicated, incubated for 5 min and imaged immediately upon addition to the coverslip. Images were acquired with a Nikon TiE equipped with a 60 \times plan-apochromat 1.2 numerical aperture W objective and Yokagawa CSU-X1 scan head. Images were acquired on an Andor DU-897 back-illuminated CCD. Acquired images were processed by the SQUASH package for Fiji³⁷.

Confocal microscopy of bead-supported membrane tethering assay. A 200 μ l observation chamber (μ -Slide 8 well, uncoated, #1.5, ibidi) was pre-blocked with BSA (1 mg/ml in standard buffer) for 1.5–2 h and washed thoroughly. Finally, 180 μ l of standard buffer containing beads was added to the sample chamber. In these experiments, the following concentrations were used: 1 nM Rabex-5 (131–394), 100 nM GFP-Rab5-6 \times His, and the given EEA1 concentrations (between 30 and 400 nM). Nucleotide control experiments were performed at 190 nM EEA1. Experiments were performed in standard buffer with 2 mM MgCl₂ and 1 mM nucleotide. Altogether Rab5, Rabex5, nucleotide, EEA1 and buffer were mixed in low-binding tubes at concentrations indicated, and were added to 240 μ l final volume to assure mixing throughout the chamber volume.

Images for co-localization analysis were acquired with a Nikon TiE equipped with a 60 \times plan-apochromat 1.2 numerical aperture W objective and Yokagawa CSU-X1 scan head. Images were acquired on an Andor DU-897 back-illuminated CCD. Acquired images were processed by the SQUASH package for Fiji³⁷.

Data obtained for distance measurements were acquired in the same way and processed in Fiji by determining line profiles eight pixels wide from the centre of the bead outwards over an observed vesicle. These profiles were fitted with a Gaussian distribution. The alignment of the microscope was confirmed by imaging of sub-diffraction beads, revealing no clear systematic shift and a maximum positional error of 21 nm determined in Motion Tracking¹⁶. Controls with sub-diffraction-sized multicolour particles (Methods) and distance measurements between Rab5

itself and its resident membrane were within the measurement error of the technique (approximately 15 nm)³⁸.

Super-resolution imaging of EEA1 termini. HeLa cells were stained using primary antibodies against EEA1 N terminus (610457, prepared in mouse, BD Biosciences) and EEA1 C terminus (2900, prepared in rabbit, Abcam). The secondary antibodies were anti-mouse Alexa568 antibody (A-11004, prepared in goat, Life Technologies) and anti-rabbit Alexa647 (A-21244, prepared in goat, Life Technologies). Coverslips were mounted in STORM buffer (100 mM Tris-HCl pH8.7, 10 mM NaCl, 10% glucose, 15% glycerol, 0.5 mg/ml glucose oxidase, 40 µg/ml catalase, 1% BME) and sealed with nail polish. Cells were imaged on a Zeiss Eclipse Ti microscope equipped with a 150 mW 561 nm laser and a 300 mW 647 laser. For imaging, lasers intensities were set to achieve 50 mW at the rear lens of the objective. Illumination was applied at a sub-TIRF angle through the objective to improve the signal to noise ratio. Videos of 24,000 frames (12,000 frames per channel) were acquired by groups of 6 consecutive frames using the NIS Elements software (Nikon). Images were aligned using 100 nm Tetraspeck beads (Thermo Fisher). This software was also used for peak detection and image reconstruction. The localization of the EEA1 termini could be distorted a maximum of approximately 20 nm owing to the size of the antibodies. The localization accuracy of the secondary antibody was ~25 nm. Measured distances were determined in Fiji and represent distances between respective centres-of-mass. Representative experiment is shown, $n = 3$.

Sample preparation for optical trap experiments. Bead-supported membranes were prepared as described. The concentrations used were as in the microscopy experiments: 1 nM Rabex-5 (131–394), 100 nM Rab5-6×His and EEA1 concentrations (between 30 and 400 nM). Most experiments were performed at 40 nM EEA1, with additional trials taking place at 4 and 400 nM. At lowest concentrations, single transient events became difficult to observe (<5% had interactions). At the highest concentrations, events were often non-transient or repeated.

Electron microscopy. Samples were rotary-shadowed essentially as described³⁹. Briefly, samples were diluted in a spraying buffer, consisting of 100 mM ammonium acetate and 30% glycerol. Diluted samples were sprayed via a capillary onto freshly cleaved mica chips. These mica chips were mounted in the high vacuum evaporator (MED 020, Baltec) and dried. Specimens were platinum coated (5–7.5 nm) and carbon was evaporated. Following deposition, the replica was floated off and examined at 71,000× magnification and imaged onto a CCD (Morgagni 268D, FEI; Morada G2, Olympus).

Analysis of electron microscopy. Images obtained were processed in ImageJ by skeletonizing the particles. Lengths were determined directly from these data and represent an overestimation due to the granularity of the platinum shadowing (5–7.5 nm granules). The bouquet plots were generated by aligning the initial five segments of the molecules and the entire population set was plotted.

To determine the curvature measure, we first took the skeletonized curves and smoothed them with a window of 8.2 nm. These curves were then segmented with 301 equally spaced points, and these smoothed curves were used for the curvature calculation. We first attempted to define curvature at one segment length (~0.75 nm) but this analysis was too noisy to obtain meaningful description of the curves. We therefore determined the curvature by taking the difference of the tangents and dividing it by the arc length at a distance of ~15 nm (20 points). The variance of this measure was determined, and bootstrapping with resampling was used to determine errors over the whole population and for 1,000 iterations.

Although proteins are not homogeneous polymers, the WLC model captures essential aspects of the physics underlying their shape fluctuations^{40,41}. Calculation of fits to all mean tangent-correlations and the equilibration analysis were performed using Easyworm source code in Matlab⁴². First, the original skeletonized curves were segmented with 301 equally spaced points. These data were then used to calculate the tangent-correlations and the kurtosis plots. We fitted the regime whereby the kurtosis measurement defined that the molecules were equilibrated^{18,43,44}. This distance therefore varied (see Extended Data Fig. 6, kurtosis plots), but the estimation of persistence length was only weakly dependent on this distance. The fitting routines were then implemented up to the thermal equilibration distance with bootstrapping with resampling, which was run for the whole population and 1,000 times to obtain errors. These are given as mean ± standard deviation. For values and fit statistics, please refer to Supplementary Data Table. We did not apply the WLC model to the swapped mutant (Extended Data Fig. 4h) because of the lack of significant structural changes upon Rab5 binding (Fig. 2f and Extended Data Fig. 4f).

The analytical fitting to the radial distribution functions was performed in Python¹⁸. The radial distribution function for a worm-like chain is the probability density for finding the end points of the polymer. The polymers are considered as embedded in a two-dimensional space in this scheme. This treatment adopts the continuum model of the polymer, thereby defining the statistical properties via free energy calculation. Fitting to analytical solution of the WLC yielded a

mean effective persistence length of 270 ± 14 nm for EEA1 alone (mean ± error of fit), and two populations of effective persistence lengths (26 ± 2 nm (67%) and 300 ± 14 nm (33%)) for EEA1 in the presence of Rab5-GTP-γS.

Optical tweezer experiments. A custom-built high-resolution dual-trap optical tweezer microscope was used^{45,46}. A single stable solid-state laser (Spectra-Physics, 5 W) was split by polarization into two traps that could be independently manoeuvred. Forces were measured independently in both traps by back-focal plane interferometry. Absolute distances between the two traps were determined by template-based video microscopy analysis (43 ± 2 nm per pixel) and offset-corrected for each microsphere pair by repeatedly contacting the microspheres after each experiment. The template detection algorithm had subpixel accuracy, at an estimated uncertainty in absolute distance measurements to be not more than ±20 nm. Bead displacement was calculated according to $\Delta F = -\kappa\Delta y$. Extended Data Fig. 7g demonstrates the sensitivity of the instrument via the Allan deviation⁴⁷ for averaging times greater than 100 ms.

All optical tweezer experiments were performed with 2 µm silica size-standard microspheres (Corpuscular), at a temperature of 26 ± 2 °C in a laminar flow chamber with buffers containing 35% glycerol to prevent sedimentation of the silica microspheres. Thermal calibration of the optical traps was performed with the power spectrum method using a dynamic viscosity of 3.1 mPas (ref. 48) (mean trap stiffness: trap 1, $\kappa_1 = 0.035 \pm 0.007$ pN/nm; trap 2, $\kappa_2 = 0.029 \pm 0.007$ pN/nm), leading to an overall trap stiffness of $\kappa_T = 0.0159$ pN/nm (yellow response curve in Extended Data Fig. 7h). Data were acquired at 1 kHz and further processed using custom-written software in R. Spurious electronic noise at 50 Hz was filtered using a fifth-order Butterworth notch filter from 49 to 51 Hz.

For probing the interactions of EEA1 with Rab5 without any assumptions on the shape of EEA1, a distance agnostic protocol with consecutive cycles of approaching, waiting (20 s) and retraction was used, approaching closer in each iteration (Fig. 3b). The stationary segments were then subjected to automatic change-point analysis to identify regions of the time series longer than 100 ms with significantly different mean and variance⁴⁹. Events thus identified were classified as transient if the mean and variance went back to base levels within the stationary segment (see examples in force traces in Fig. 3c and Extended Data Fig. 7). Mean times of interactions were 3.4 ± 0.6 s for GTP-γS and 0.9 ± 0.2 s for GTP. A fluctuation analysis of the differential distance signal during these events gave an estimated tether misalignment of less than 30° in all interactions⁵⁰. Only transient events were further processed. Silica beads alone as a negative control measured a mean contact distance of 22 nm (Fig. 3d, grey).

To calculate the persistence length for individual captured molecules we determined the equilibrium extension, z_{eq} , from the capture distance D (nm), the average measured force increase upon tethering ΔF (pN) and the known displacements from each trap $\Delta x_1 = \Delta F/\kappa_1$ and $\Delta x_2 = \Delta F/\kappa_2$ as $z_{eq} = D - \Delta x_1 - \Delta x_2$. With this distance, the persistence length was calculated according to⁵¹

$$\lambda(\Delta F, z_{eq}) = \frac{k_B T}{\Delta F} \left(\frac{z_{eq}}{L} - \frac{1}{4} + \frac{1}{4(1 - z_{eq}/L)^2} \right)$$

Similarly, to estimate the magnitude of the entropic collapse force, this formula was applied to the equilibrium extensions of EEA1, as estimated by the end-to-end distances of the molecules from electron microscopy. Values determined were (median and bounds at (2.5%, 97.5%)) EEA1, 23 (14, 33) nm; extended, 73 (60, 88) nm; swapped, 26 (21, 30) nm; 10×His, 78 (35, 140) nm. Values reported are medians and 95% confidence intervals determined from bootstrapping.

Generation of HeLa EEA1-KO cell line. HeLa EEA1-KO lines were generated using CRISPR-Cas9 technology⁵² on HeLa-Kyoto cell lines obtained from the BAC recombineering facility at the Max Planck Institute of Molecular Cell Biology and Genetics. Cell lines were tested for mycoplasma and authenticated (Multiplexion, Heidelberg). pSpCas9(BB-2A-GFP (PX458) and pSpCas9(BB)-2A-Puro (PX459) were a gift from F. Zhang (Addgene plasmid 48138, 48139). A PX458 plasmid encoding a GFP-labelled Cas9 nuclease and the sgRNA sequence (from GECKO⁵² library 17446, GTGGTTAAACCATGTTAAGG, targeting first exon) was transfected into standard HeLa Kyoto cells with Lipofectamine 2000 following the manufacturer's instructions. Cells were cultured in DMEM media supplemented with 10% FBS and 1% penicillin-streptomycin at 37 °C and 5% CO₂. After 3 days, the transfected cells were FACS sorted by their GFP fluorescence into 96-well plates to obtain single clones and visually inspected⁵³. These clones were then screened by western blotting and in-del formation confirmed sequencing of genomic DNA (primer forward, AGCGCCGTCGCCACCG; reverse, TAAGCGCCTGCCGGGCTG). Note the region is extremely GC-rich (75%, ±250 nt from targeted indel region). Additionally, a mixed-clonal line was obtained by transfection of HeLa Kyoto with PX459 with the above sgRNA sequence. After 72 h from transfection, cells were exchanged into media supplemented with 0.5 µg/ml puromycin (concentration determined in separated

experiment) and selected for 3 days. All imaging experiments were confirmed on this secondary line.

Endocytosis rescue assays. Wild-type EEA1 and the extended and swapped variants (Extended Data Fig. 3) were cloned into customized mammalian expression plasmids under the CMV promoter resulting in untagged proteins. HeLa or HeLa EEA1-KO cells were seeded into 96-well plates and transfected (or mock transfected) after 48 h. Following 48 h after transfection, cells were exchanged into serum-free media containing 8.2 µg/ml LDL-Alexa 488 (prepared as previously described¹⁶) or 100 ng/ml EGF-Alexa 488 (E13345, Thermo Fisher) for 10 min at 37 °C, and washed in PBS then fixed in 4% paraformaldehyde.

Automated confocal immunofluorescence microscopy and analysis. Fixed cells were stained with antibodies against EEA1 (laboratory-made rabbit) and Rab5 (610724, prepared in mouse, BD Biosciences) as previously described²⁴. DAPI was used to stain the nuclei. Not all early endosomes harbour EEA1 (ref. 54) and other tethering factors could compensate for EEA1 (refs 24, 55). All imaging was performed on a Yokogawa CV7000 s automated spinning disc confocal using a 60 × 1.2 numerical aperture objective. Fifteen images were acquired per well and each condition was duplicated at least twice per plate, resulting in 30 or more images per condition.

Image analysis used home-made software, MotionTracking, as previously described^{56,57}. Images were first corrected for illumination, chromatic aberration and physical shift using multicolour beads. All cells, nuclei and cell objects in corrected images were then segmented and their size, content and complexity calculated. The intensity of EEA1 in wild-type HeLa cells was measured to determine a wild-type intensity distribution. In the rescue experiments, an intensity threshold for the transfections was set at about two times the mean of wild-type cells (Extended Data Fig. 8i). Experiments were repeated at different seeding densities with similar results. Given a cell density threshold between 10 and 100 per image, we obtained an average of more than 300 cells per condition after filtering for the transfection level of EEA1, and more than 15,000 endosomes per experiment. A two-tailed *t*-test was used for significance calculations.

Cell electron microscopy. Cells in 3 cm diameter plastic dishes were processed for electron microscopy using a method⁵⁸ to provide particularly heavy staining of cellular components. Briefly, cells were fixed by addition of 2.5% glutaraldehyde in PBS for 1 h at room temperature and then washed with PBS. The cells were then processed as described⁵⁸ with sequential incubations in solutions containing potassium ferricyanide/osmium tetroxide, thiocarbonylhydrazide, osmium tetroxide, uranyl acetate and lead nitrate in aspartic acid before dehydration and flat embedding in resin. Sections were cut parallel to the substratum and analysed unstained in a JEOL 1011 transmission electron microscope (Tokyo, Japan). Images for quantitation were collected from coded samples (double blind) to avoid bias.

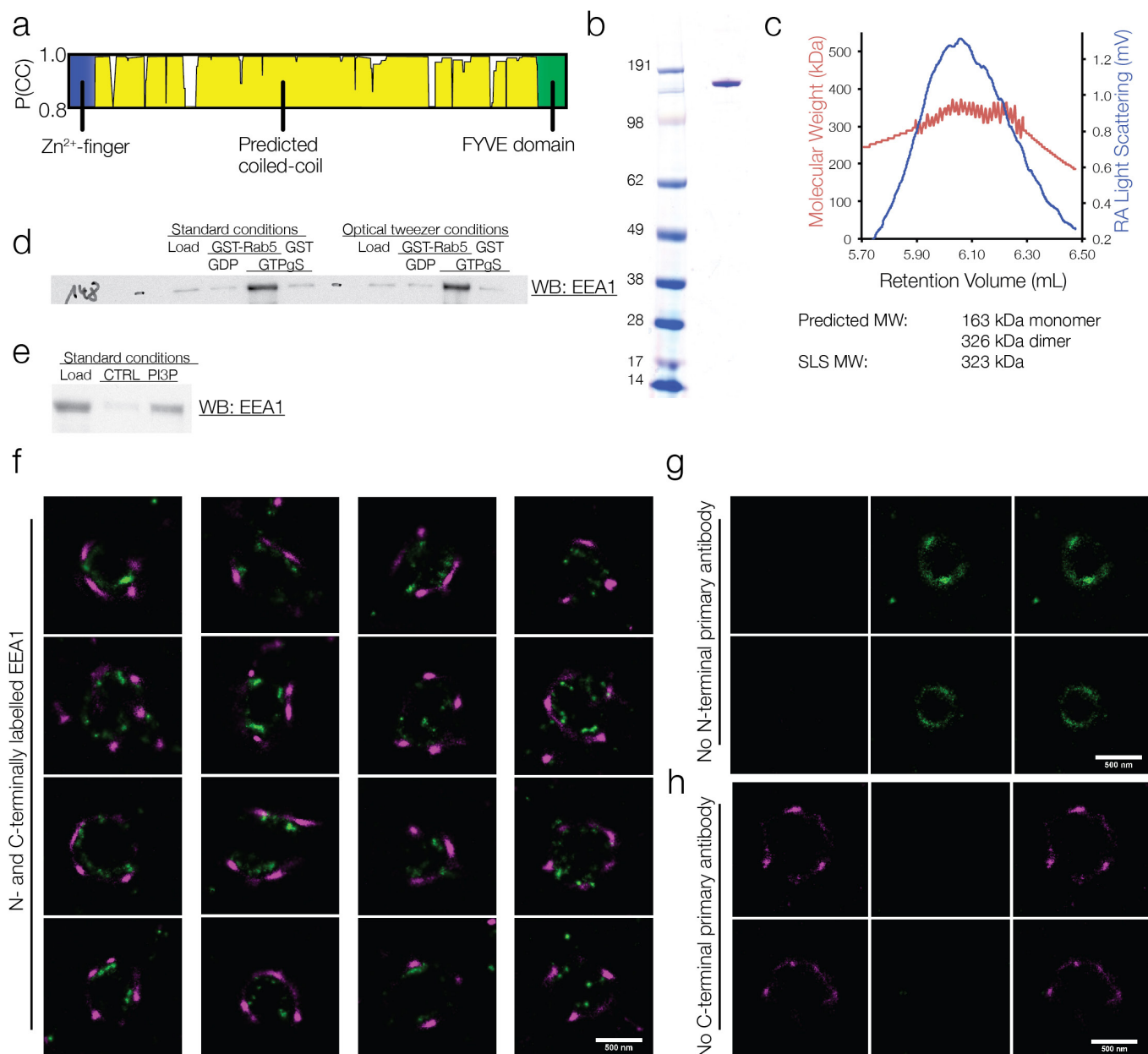
Distance analysis used ImageJ. To correct for thickness of slices (60 nm), the following equation was used:

$$P(R) = \frac{1}{Z} \int_0^H P_0(\sqrt{R^2 - h^2}) \frac{R}{\sqrt{R^2 - h^2}} dh,$$

where $P_0(r)$ is the apparent 2D distance distribution, R is the 3D distance, H is the thickness of the slice and Z is the normalization constant. Uncorrected distance was measured at 119.8 ± 78.2 nm (mean \pm s.d.), which resulted in 130.0 ± 76.8 nm corrected.

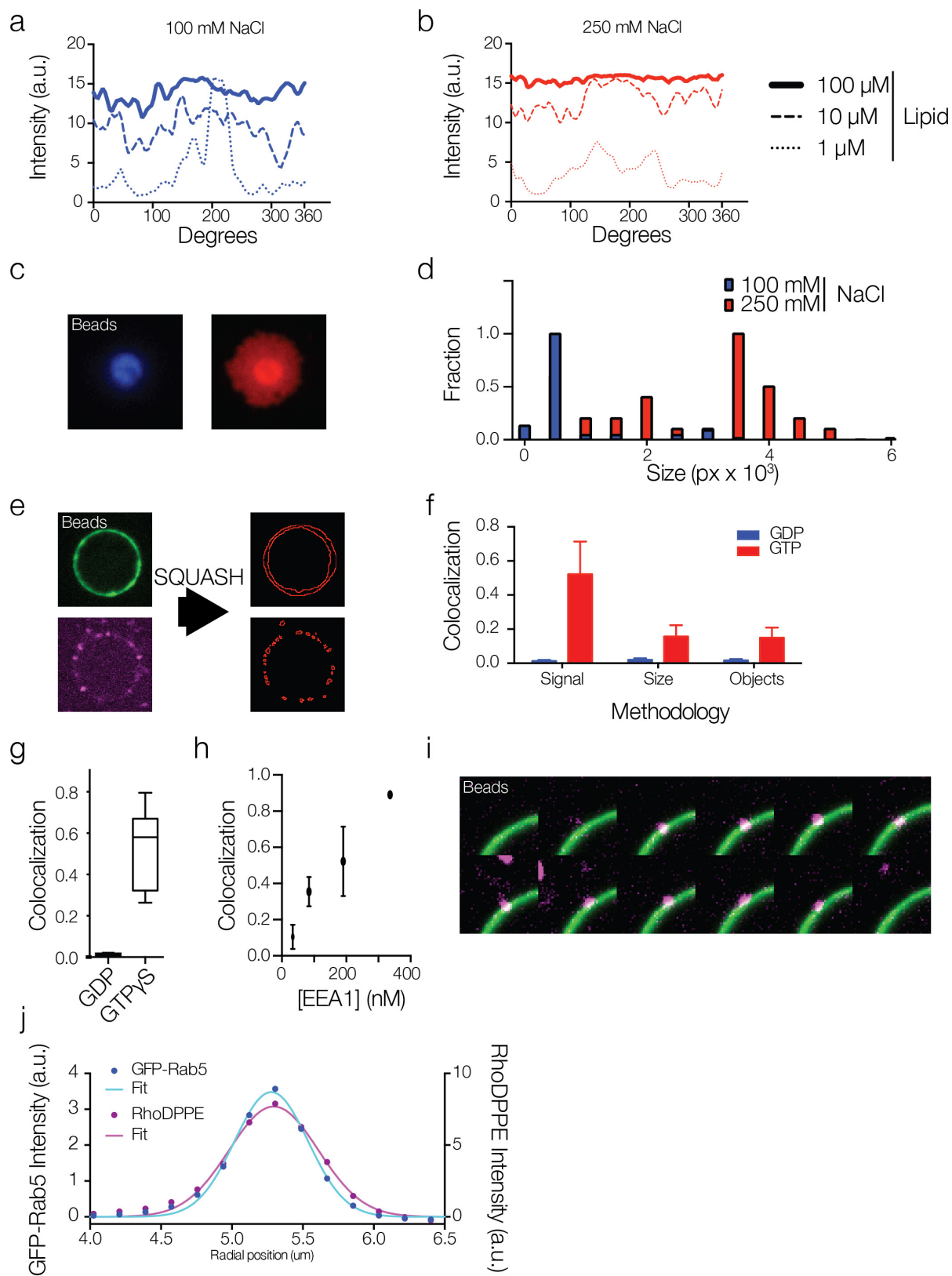
30. Sheffield, P., Garrard, S. & Derewenda, Z. Overcoming expression and purification problems of RhoGDI using a family of “parallel” expression vectors. *Protein Expr. Purif.* **15**, 34–39 (1999).
31. Delprato, A., Merithew, E. & Lambright, D. G. Structure, exchange determinants, and family-wide rab specificity of the tandem helical bundle and Vps9 domains of Rabex-5. *Cell* **118**, 607–617 (2004).

32. Horiuchi, H. *et al.* A novel Rab5 GDP/GTP exchange factor complexed to Rabaptin-5 links nucleotide exchange to effector recruitment and function. *Cell* **90**, 1149–1159 (1997).
33. Boura, E. & Hurley, J. H. Structural basis for membrane targeting by the MVB12-associated β -prism domain of the human ESCRT-I MVB12 subunit. *Proc. Natl Acad. Sci. USA* **109**, 1901–1906 (2012).
34. Murray, D. H., Tamm, L. K. & Kiessling, V. Supported double membranes. *J. Struct. Biol.* **168**, 183–189 (2009).
35. Neumann, S., Pucadyil, T. J. & Schmid, S. L. Analyzing membrane remodeling and fission using supported bilayers with excess membrane reservoir. *Nature Protocols* **8**, 213–222 (2013).
36. Pucadyil, T. J. & Schmid, S. L. Real-time visualization of dynamin-catalyzed membrane fission and vesicle release. *Cell* **135**, 1263–1275 (2008).
37. Rizk, A. *et al.* Segmentation and quantification of subcellular structures in fluorescence microscopy images using Squash. *Nature Protocols* **9**, 586–596 (2014).
38. Lo, S. Y. *et al.* Intrinsic tethering activity of endosomal Rab proteins. *Nature Struct. Mol. Biol.* **19**, 40–47 (2011).
39. Tyler, J. M. & Branton, D. Rotary shadowing of extended molecules dried from glycerol. *J. Ultrastruct. Res.* **71**, 95–102 (1980).
40. Gittes, F., Mickey, B., Nettleton, J. & Howard, J. Flexural rigidity of microtubules and actin filaments measured from thermal fluctuations in shape. *J. Cell Biol.* **120**, 923–934 (1993).
41. Eeftens, J. M. *et al.* Condensin Smc2-Smc4 dimers are flexible and dynamic. *Cell Reports* **14**, 1813–1818 (2016).
42. Lamour, G., Kirkegaard, J. B., Li, H., Knowles, T. P. & Gsponer, J. Easyworm: an open-source software tool to determine the mechanical properties of worm-like chains. *Source Code Biol. Med.* **9**, 16 (2014).
43. Rivetti, C., Guthold, M. & Bustamante, C. Scanning force microscopy of DNA deposited onto mica: equilibration versus kinetic trapping studied by statistical polymer chain analysis. *J. Mol. Biol.* **264**, 919–932 (1996).
44. Valle, F., Favre, M., De Los Rios, P., Rosa, A. & Dietler, G. Scaling exponents and probability distributions of DNA end-to-end distance. *Phys. Rev. Lett.* **95**, 158105 (2005).
45. Lisica, A. *et al.* Mechanisms of backtrack recovery by RNA polymerases I and II. *Proc. Natl Acad. Sci. USA* **113**, 2946–2951 (2016).
46. Jahnel, M., Behrndt, M., Jannasch, A., Schäffer, E. & Grill, S. W. Measuring the complete force field of an optical trap. *Opt. Lett.* **36**, 1260–1262 (2011).
47. Czerwinski, F., Richardson, A. C. & Oddershede, L. B. Quantifying noise in optical tweezers by allan variance. *Opt. Express* **17**, 13255–13269 (2009).
48. Nørrelykke, S. F. & Flyvbjerg, H. Power spectrum analysis with least-squares fitting: amplitude bias and its elimination, with application to optical tweezers and atomic force microscope cantilevers. *Rev. Sci. Instrum.* **81**, 075103 (2010).
49. Killick, R., Fearnhead, P. & Eckley, I. A. Optimal detection of changepoints with a linear computational cost. *J. Am. Stat. Assoc.* **107**, 1590–1598 (2012).
50. Ribezzi-Crivellari, M. & Ritort, F. Force spectroscopy with dual-trap optical tweezers: molecular stiffness measurements and coupled fluctuations analysis. *Biophys. J.* **103**, 1919–1928 (2012).
51. Marko, J. F. & Siggia, E. D. Statistical mechanics of supercoiled DNA. *Phys. Rev. E* **52**, 2912–2938 (1995).
52. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8**, 2281–2308 (2013).
53. Poser, I. *et al.* BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nature Methods* **5**, 409–415 (2008).
54. Kalaidzidis, I. *et al.* APPL endosomes are not obligatory endocytic intermediates but act as stable cargo-sorting compartments. *J. Cell Biol.* **211**, 123–144 (2015).
55. Peplowska, K., Markgraf, D. F., Ostrowicz, C. W., Bange, G. & Ungermann, C. The CORVET tethering complex interacts with the yeast Rab5 homolog Vps21 and is involved in endo-lysosomal biogenesis. *Dev. Cell* **12**, 739–750 (2007).
56. Collinet, C. *et al.* Systems survey of endocytosis by multiparametric image analysis. *Nature* **464**, 243–249 (2010).
57. Gilleron, J. *et al.* Image-based analysis of lipid nanoparticle-mediated siRNA delivery, intracellular trafficking and endosomal escape. *Nature Biotechnol.* **31**, 638–646 (2013).
58. Takasato, M. *et al.* Kidney organoids from human iPS cells contain multiple lineages and model human nephrogenesis. *Nature* **526**, 564–568 (2015).



Extended Data Figure 1 | EEA1 is a predicted extended coiled-coil dimer that binds Rab5 in a GTP-dependent manner and extends outwards from endosomes **a**, Human EEA1 in COILS prediction reveals a clear coiled-structure flanked by the Rab5-binding Zn²⁺-finger on the N terminus and PI(3)P binding FYVE domain on the C terminus. **b**, Coomassie-stained gel of human EEA1 expressed as a GST fusion in SF+ insect cells and purified by GS affinity, cleaved on resin, and subsequently concentrated and separated from smaller contaminants by size-exclusion chromatography on a Superose 6 column. **c**, Static light scattering in line with size-exclusion chromatography reveals a molecular mass of 323 kDa, compared with a theoretical molecular mass of 326 kDa for a dimeric protein. **d**, Purified protein binds Rab5 in both standard and optical tweezer conditions (35% glycerol) in a GTP-dependent manner. GST or GST-Rab5 was purified and conjugated to GS resin, and subsequently nucleotide was exchanged to either GTP- γ S or GDP using

EDTA-Mg²⁺-mediated exchange and subsequent wash. The GST resin was then incubated with EEA1 in either the standard or optical tweezers buffer, washed three times, and beads were then blotted for EEA1. **e**, Recombinant EEA1 binds specifically to PI(3)P liposomes. When mixed with POPC:POPS 85:15 liposomes, no EEA1 is observed in the liposome pellet (CTRL). In contrast, EEA1 is pelleted with control POPC:POPS:PI(3)P 80:15:5 liposomes (PI3P). **f**, The N-terminal Zn²⁺-finger and C-terminal FYVE domain of EEA1 were differentially labelled with specific antibodies and STORM microscopy performed to define their localization in HeLa cells. Representative STORM images of EEA1 radial extension from endosome of $n = 22$. Scale bar, 500 nm. **g, h**, Primary antibody binding controls for N and C termini. Primary antibodies for the N (**g**) and C (**h**) termini were left out of the staining, resulting in no unspecific secondary staining for each. Representative of $n = 5$. Scale bar, 500 nm.



Extended Data Figure 2 | See next page for caption.

Extended Data Figure 2 | Validation of bead-supported lipid bilayers for optical tweezers, and bead tethering experiment controls and methods.

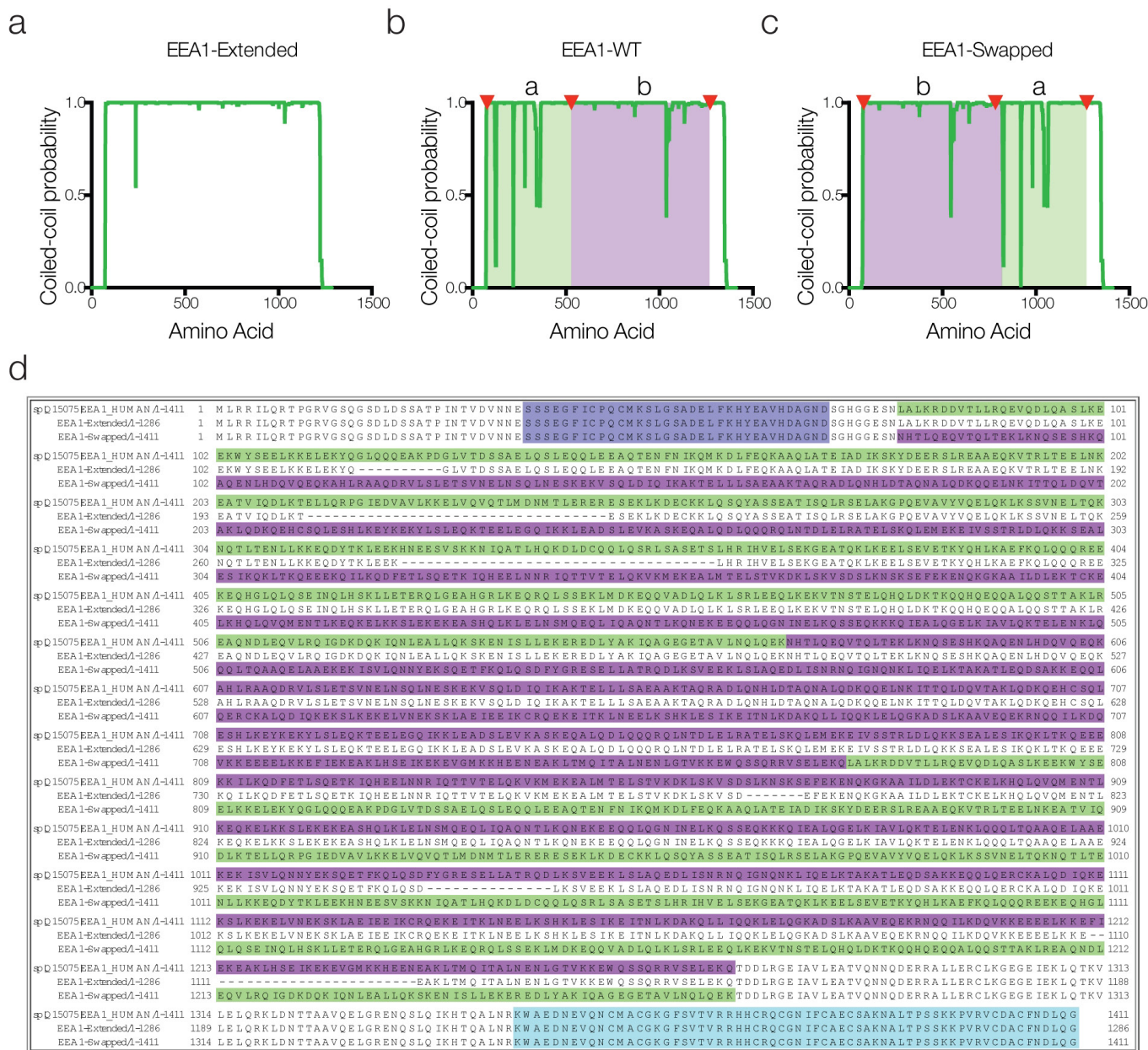
To optimize the conditions for forming supported lipid bilayers on the 2–10 μm beads, we systematically investigated the dependence of membrane formation on salt and liposome concentration. **a**, Fluorescent profiles of supported lipid bilayer bead cross sections. At high liposome concentration (100 μM , solid line) during formation of the bilayer on the silica bead, the bead-supported membrane fluorescence intensity is circumferentially homogenous. At lower lipid concentrations (10 and 1 μM , dashed and dotted lines), less than full coverage is achieved and the supported bilayer is inhomogeneous. **b**, Consistent with previous reports, increasing salt concentrations result in more homogenous membrane coverage. **c**, Representative examples of the ‘spilled-out’ membrane of beads prepared at 100 mM (top, blue) and 250 mM (bottom, red) NaCl salt and 100 μM liposomes, of $n = 5$. **d**, Histogram of the size of membrane spilled from the beads onto the substrate when prepared at 100 and 250 mM NaCl (blue and red, respectively). This indicated that the lower salt samples (blue) were homogeneously covered with membrane and that they had little excess present, and therefore the optimal conditions for formation of membrane on the silica beads used in tethering and in optical tweezer experiments. **e**, Segmentation of beads and vesicles by

the SQUASH method. Bead-supported bilayers and vesicles (green and magenta, respectively) were segmented as illustrated by red outlines to determine their co-localization. Representative of $n = 1$ generated for schematic. **f**, Methodology comparison for co-localization in GDP and GTP- γS conditions. All methods give $P < 0.01$ in a two-tailed Student's t -test. Co-localization by signal is better than by size or object, as vesicles become undercounted at high concentrations. Mean \pm s.d., $n = 5$.

g, Co-localization of liposomes (PI(3)P, magenta) to the bead-supported membrane (GFP-Rab5, green) was strictly dependent on GTP- γS .

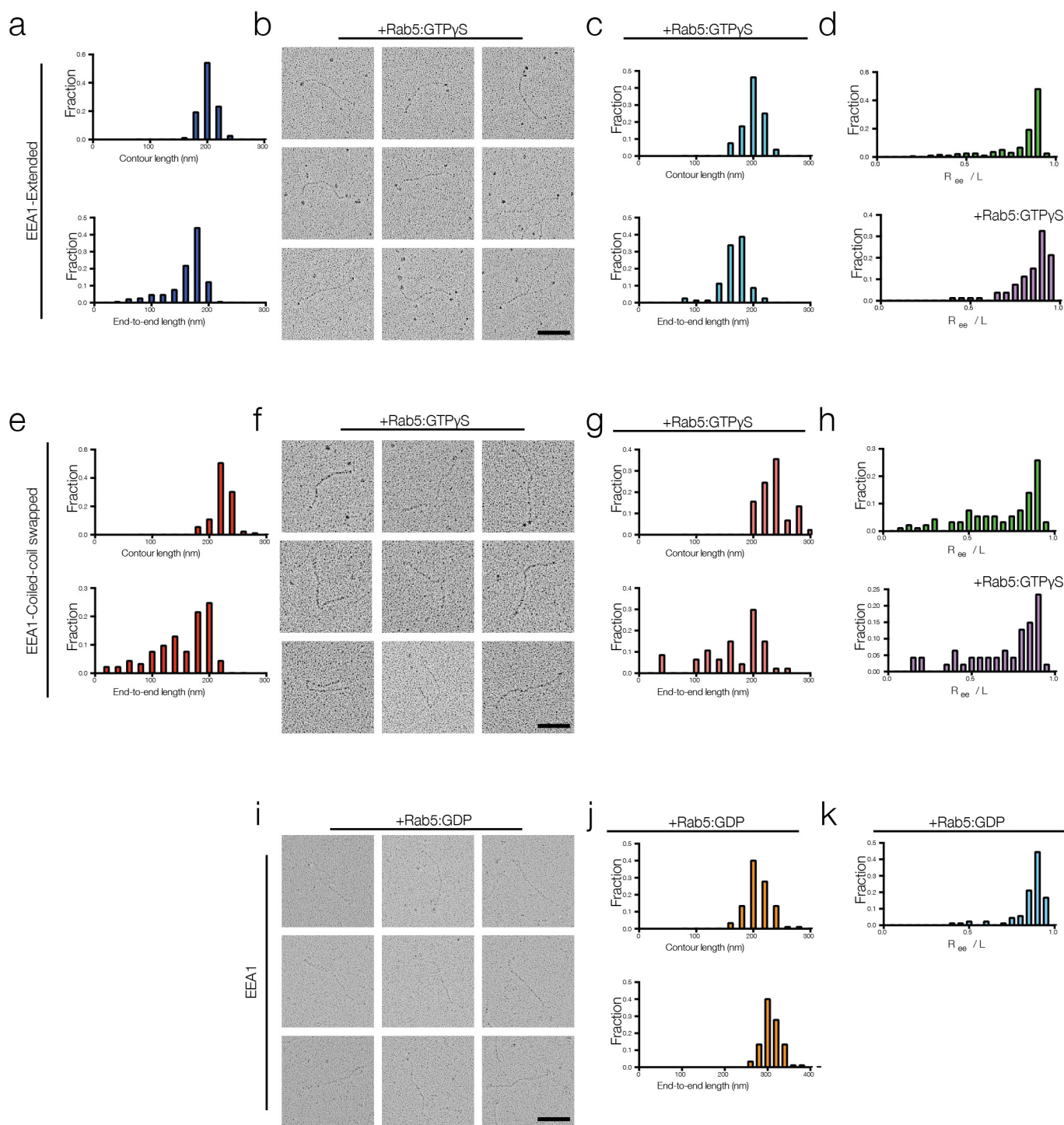
Box-whisker plot with minimum/maximum error, $n = 5$. **h**, The co-localization of liposomes to the supported membrane was dependent on EEA1 concentration. At higher concentrations of EEA1, co-localization approached 100%. These concentrations are within the range of the concentration of endogenous protein²³. Mean \pm s.d., $n = 5$. **i**, Time-lapse micrographs of the bead-supported bilayer labelled with GFP-Rab5 (green), and a dynamically tethered vesicle (magenta). Vesicles were observed to tether and reversibly leave the membrane, as well as diffuse about its surface. Images displayed were acquired at 350 ms intervals as z -stacks. Representative of $n = 1$ to acquire video. Scale bar, 2 μm .

j, Example fits for radial line-profile data.



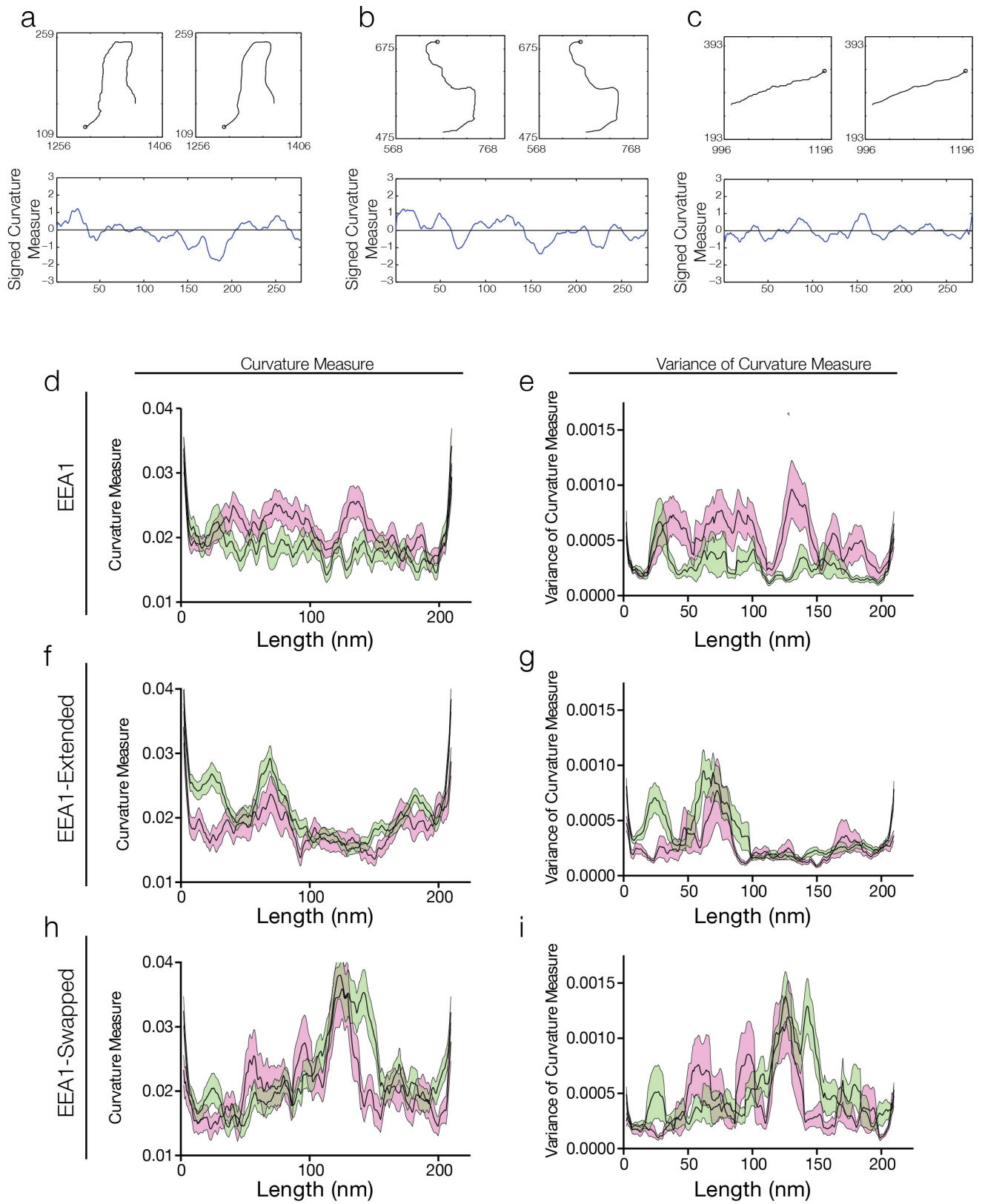
Extended Data Figure 3 | Structure prediction and sequence description of EEA1 mutants. a, COILS prediction for extended EEA1 mutant, revealing removal of most of the discontinuities in the coiled-coil. **b, c**, The swapped EEA1 mutant has a rearranged coiled-coil. The coiled-coil was split as indicated by red triangles in the original EEA1-WT (**b**), and the two regions *a* (shaded green) and *b* (shaded magenta) were rearranged in a synthetic gene, producing the swapped EEA1 variant maintaining the features and sequence of the original coiled-coil, but in an alternative

location (**c**). **d**, Full sequence alignment for human EEA1 and the extended and swapped mutants used in the study. The crystal structure (Protein Data Bank accession number 3MJH) for the Zn²⁺-finger domain is marked in dark blue close to the N terminus. Segment *a* of the coiled-coil region is marked in green, and segment *b* in magenta. The crystal structure (Protein Data Bank accession number 1JOC) of the C-terminal FYVE domain and portion of the coiled-coil is marked in cyan. Details of the mutant constructs are found in the Methods.



Extended Data Figure 4 | Extended and swapped EEA1 mutants exhibit limited changes in the presence of Rab5:GTP- γ S. **a, e,** Rotary-shadowed EEA1-extended particles and EEA1-swapped mutants were skeletonized and analysed in ImageJ for contour length (top), resulting in normally distributed contour length histograms. The end-to-end length histograms (bottom) are similarly distributed. These data were collected on N-terminally MBP-tagged samples. Compare with wild-type in Fig. 2b, d; $n = 212$ for the extended and $n = 93$ for the swapped variants. **b–d, f, g,** The EEA1 mutants revealed limited changes to their curvature

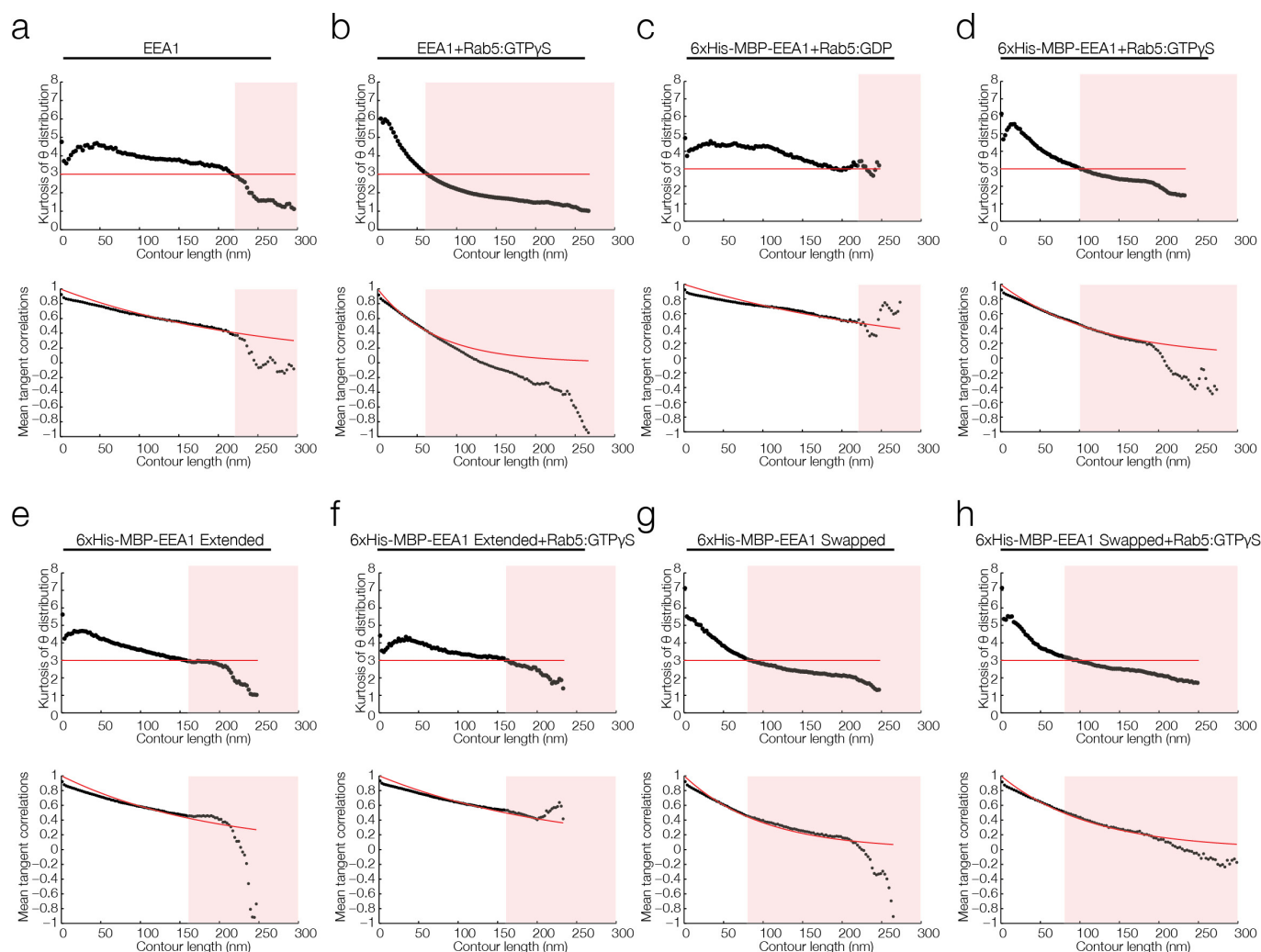
in the presence of Rab5:GTP- γ S (**b, f**; compare Fig. 2i, j), and therefore minor changes to their contour and end-to-end length histograms (**c, g**) and radial distribution plots (**d, h**); $n = 80$ for the extended and $n = 47$ for the swapped variants. **i, j,** Rotary-shadowing electron microscopy of EEA1 in the presence of Rab5:GDP ($n = 90$), N-terminally MBP-tagged, revealed no change in appearance compared with the absence of Rab5 entirely (Fig. 2a), and no effect of N-terminal tagging relative to wild-type EEA1. **k,** Radial distribution function of EEA1 in the presence of Rab5:GDP (compare **d, h**; Fig. 2g); $n = 90$.



Extended Data Figure 5 | See next page for caption.

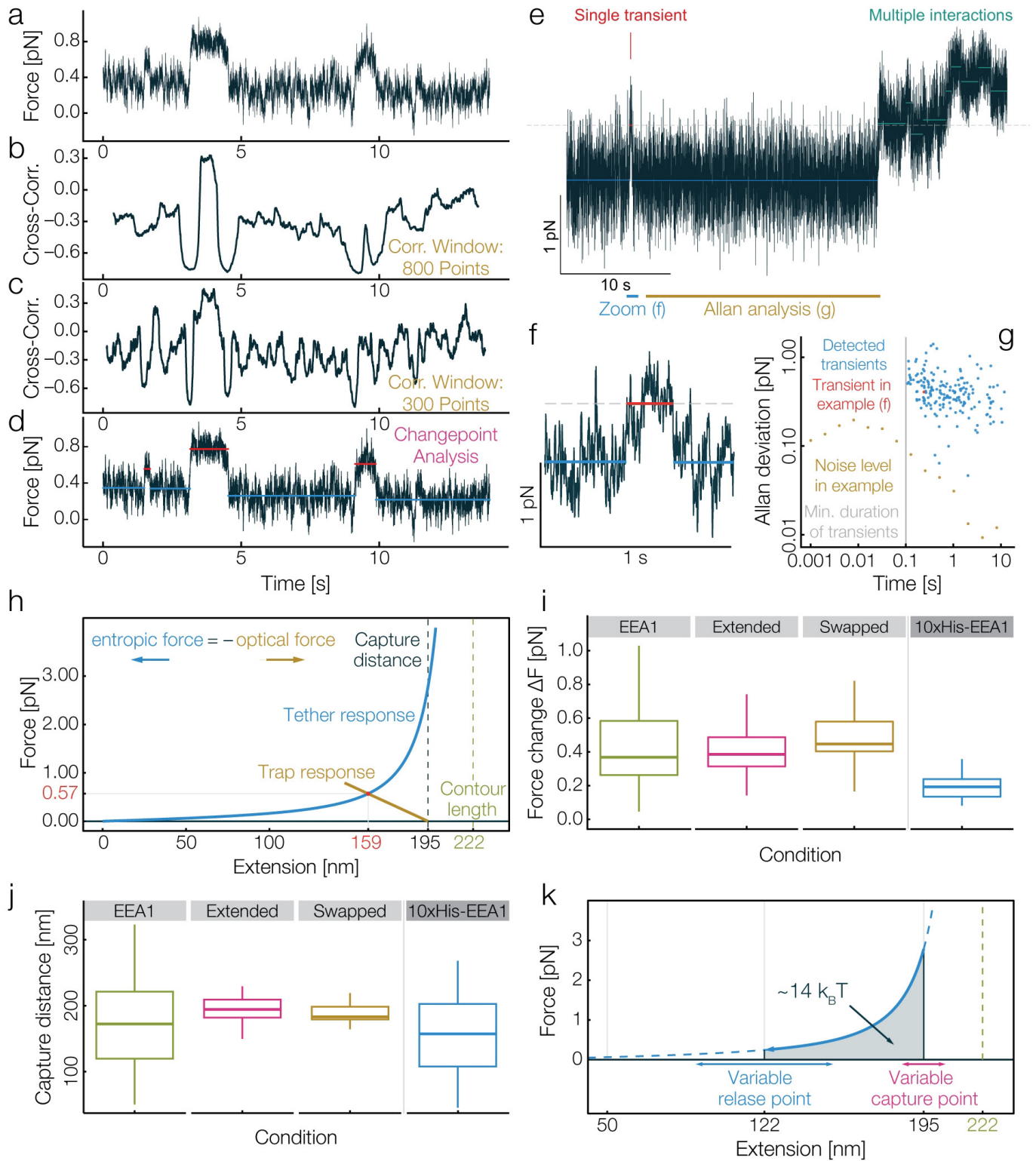
Extended Data Figure 5 | Representative segmentation, smoothing and signed curvature measures for EEA1, and averages for EEA1 and mutants. EEA1 and EEA1 mutants were skeletonized and smoothed using a moving average filter with a window of 8.2 nm, segmented to 300 equally spaced segments and aligned N terminus to C terminus by recognition of an N-terminal MBP-tag. Their curvature was calculated at 15 nm distances along the length of the proteins and plotted. **a–c**, Representative examples of rotary shadowing derived EEA1 curves. The original data appear in the first panel, with the second panel revealing the data after smoothing for comparison (Methods). The curvature measure, determined by how the tangents to the contour change at a distance of 15 nm along the contour is plotted below. Note that the choice of sign for the curvature measure is arbitrary for each molecule. **d, e**, Curvature measure and

variance of this measure for EEA1 in the presence of Rab5:GDP (green) and EEA1 in the presence of Rab5:GTP- γ S (magenta); $n = 90$, $n = 145$, respectively. Alignment of EEA1 curvature from the electron microscopy data reveals an increase in curvature over the length of the molecule upon Rab5 binding, whereas the extended and swapped EEA1 variants show no change. All curvature values were taken to be positive given that the N-terminal MBP could be recognized but the handedness of the molecule adsorbed to the grid could not be inferred. Bootstrapping with resampling at full population size was performed for 1,000 iterations to determine errors. **f, g**, Extended EEA1 variant in the absence (green) and in the presence of Rab5:GTP- γ S (magenta); $n = 212$, $n = 80$, respectively. **h, i**, Swapped EEA1 variant in the absence (green) and in the presence of Rab5:GTP- γ S (magenta); $n = 93$, $n = 47$, respectively.



Extended Data Figure 6 | Detailed persistence length and equilibration analysis for EEA1 and variants. To validate the methodology used for analysis of the persistence lengths, and to assure internal consistency in analysis methods, we systematically applied the analysis to EEA1 (and mutants, see Supplementary Data Table). The skeletonized curves were segmented to 300 equally spaced segments, where θ describes the angle between segments. The tangent–tangent correlations were then determined for the entire ensembles. **a–h**, To determine the molecular equilibration of EEA1 and variants from 3D to 2D, the kurtosis of the theta distribution (top) was calculated. Full equilibration to 2D gives a

value of 3.0, and for 3D the expected value is 1.8 as the angle distributions become Gaussian. As expected, the measured kurtosis is approximately 3.0 until lengths above the persistence length of the molecule, where the equilibration begins to fail. The value at which the kurtosis began to diverge from 2D was taken as the limit for subsequent measurements, as beyond this limit (red shaded region) 3D fluctuations are not retained and as such the consequences of surface adsorption are uncertain. Next, the tangent–tangent correlation was calculated across the ensemble and fitted up to the divergence of the kurtosis (red shaded region).

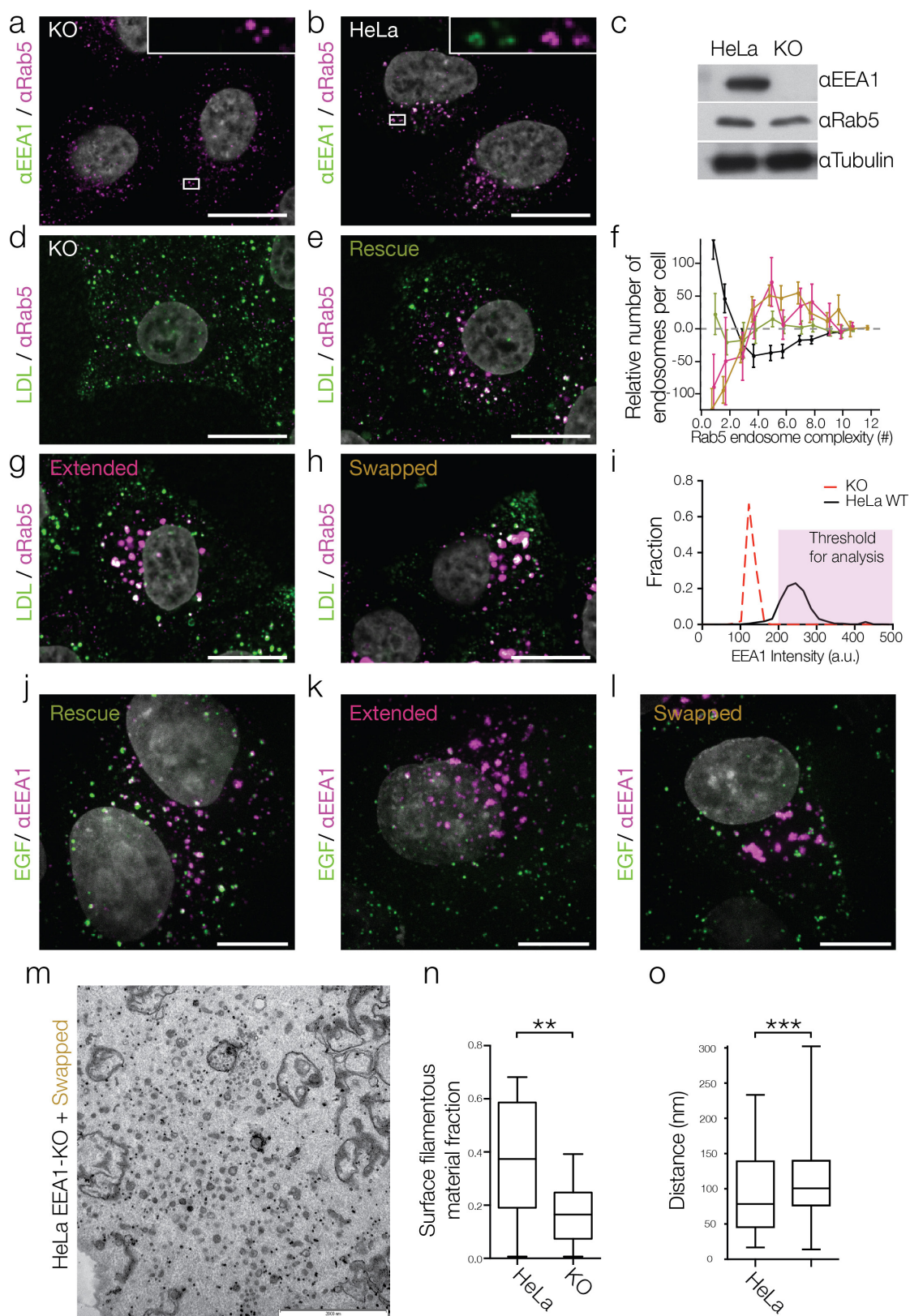


Extended Data Figure 7 | See next page for caption.

Extended Data Figure 7 | Supplementary data related to optical

tweezer experiments. a, Change-point analysis was used to identify changes in the mean and variance of the combined force signal. An example plot of averaged force (linear combination of signals from both traps) with respect to time. Data have been collected at 1 kHz. Two long transient interactions can be clearly identified. **b, c,** Cross-correlation of the force signals from each trap are not sufficient to reveal stepwise interactions as they are time-averaged. By applying cross-correlation over a correlation window of 0.8 s (**b**) or 0.3 s (**c**), long transient interactions (that is, at ~ 4 s) could be identified. However, an unbiased identification of short transients (that is, at ~ 9 s) by this method was not possible. All identified long transient interactions showed characteristic changes in the cross-correlation: anti-correlation as beads are pulled together, and correlation after tethering was established. **d,** Change-point analysis was used to detect both changes in mean and variance of the combined force signal, and thereby identify transient interactions (red line). This procedure has the additional advantage of defining clear boundaries to stepwise processes. **e,** The possibility of multiple tethers taking part in the reaction was observed. Averaged force trace for wild-type EEA1 occasionally showed signals consistent with multiple interactions (cyan), in addition to single transient interactions (red). **f,** Zoom into time series around the transient interaction identified in the previous panel. To a first approximation, the dynamic interactions were fitted as piecewise constant steps (red). Note also two very short (<10 ms) spikes of similar magnitude (to the left and right of identified interaction) occurred but are not used in further analysis. Only transients with a duration longer than 100 ms were analysed. **g,** To illustrate the sensitivity of the optical tweezer experiments, a noise analysis was performed on the segment outlined in the top panel (yellow, labelled Allan analysis). The Allan deviation (square root of Allan variance, in piconewtons) gives a threshold for detecting a signal change over different averaging windows. All detected transients (blue) are at minimum an order of magnitude above this threshold. To provide perspective, the transient in the above example is indicated as a red dot. **h,** The entropic collapse force is balanced in the tweezer experiments below its peak value. The balance between the average restoring force in the optical traps (brown) and the entropic collapse force of EEA1 (blue) in the bound state gives the measured equilibrium force

and extension (red dot). The schematic assumes the measured capture distance of 195 nm, a persistence length in the Rab5:GTP-bound state of $\lambda_b = 26$ nm, and a contour length of 222 nm. The overall trap response of the dual-trap system is treated as two springs in series with the mean trap stiffness in trap 1 ($\kappa_1 = 0.035 \pm 0.007$ pN/nm) and the mean trap stiffness in trap 2 ($\kappa_2 = 0.029 \pm 0.007$ pN/nm), leading to an overall trap stiffness of $\kappa_T = 0.0159$ pN/nm (brown line). Given these parameters, the predicted equilibrium force in the optical trap for Rab5-bound EEA1 is ~ 0.6 pN and the predicted equilibrium extension ~ 160 nm. **i,** Force changes upon capture for Rab5:GTP-bound EEA1 and the extended and swapped variants. Force was measured from change-point analysis for transient interactions between EEA1 beads and Rab5:GTP beads. To test binding per se, the force change for $10\times$ His-EEA1 beads tethered to Ni-NTA beads was similarly determined from established connections. For $10\times$ His-EEA1, no transient interactions could be observed. Median change in force and 95% confidence interval from bootstrapping with resampling (lower and upper bounds at (2.5%, 97.5%)) were determined. EEA1, 0.37 (0.31, 0.46) pN; extended, 0.39 (0.35, 0.42) pN; swapped, 0.45 (0.41, 0.56) pN; $10\times$ His, 0.19 (0.14, 0.22) pN. **j,** Capture distances defined at the proximal distance upon which transient interactions were observed for Rab5-bound EEA1 and the extended and swapped variants. Median capture distance and 95% confidence interval from bootstrapping with resampling (lower and upper bounds at (2.5%, 97.5%)) were determined. EEA1, 168 (141, 182) nm; extended, 195 (189, 199) nm; swapped, 183 (179, 189) nm; $10\times$ His, 157 (120, 196) nm; $n = 60, 93, 27, 24$ per condition respectively. **k,** Mechanical work is performed as the tether collapses. The mechanical work performed during the relaxation to the new equilibrium extension is the integral under the force–extension curve. The exact value of the extracted work depends both on the capture distance (the extension at the moment of persistence length change) and on the release distance (the extension at the moment when Rab5 unbinds). The uncertainties in these extensions are different for the two positions, reflecting the different longitudinal fluctuations of the rigid or the flexible tether ($\lambda_{\text{flexible}} = 26$ nm (blue arrows), $\lambda_{\text{rigid}} = 300$ nm (magenta arrows)). For example, for a relaxation between the capture distance, $d_{\text{capture}} \approx 195$ nm and the release extension, $d_{\text{release}} \approx 122$ nm, the extracted mechanical work is $W \approx 14 k_B T$.

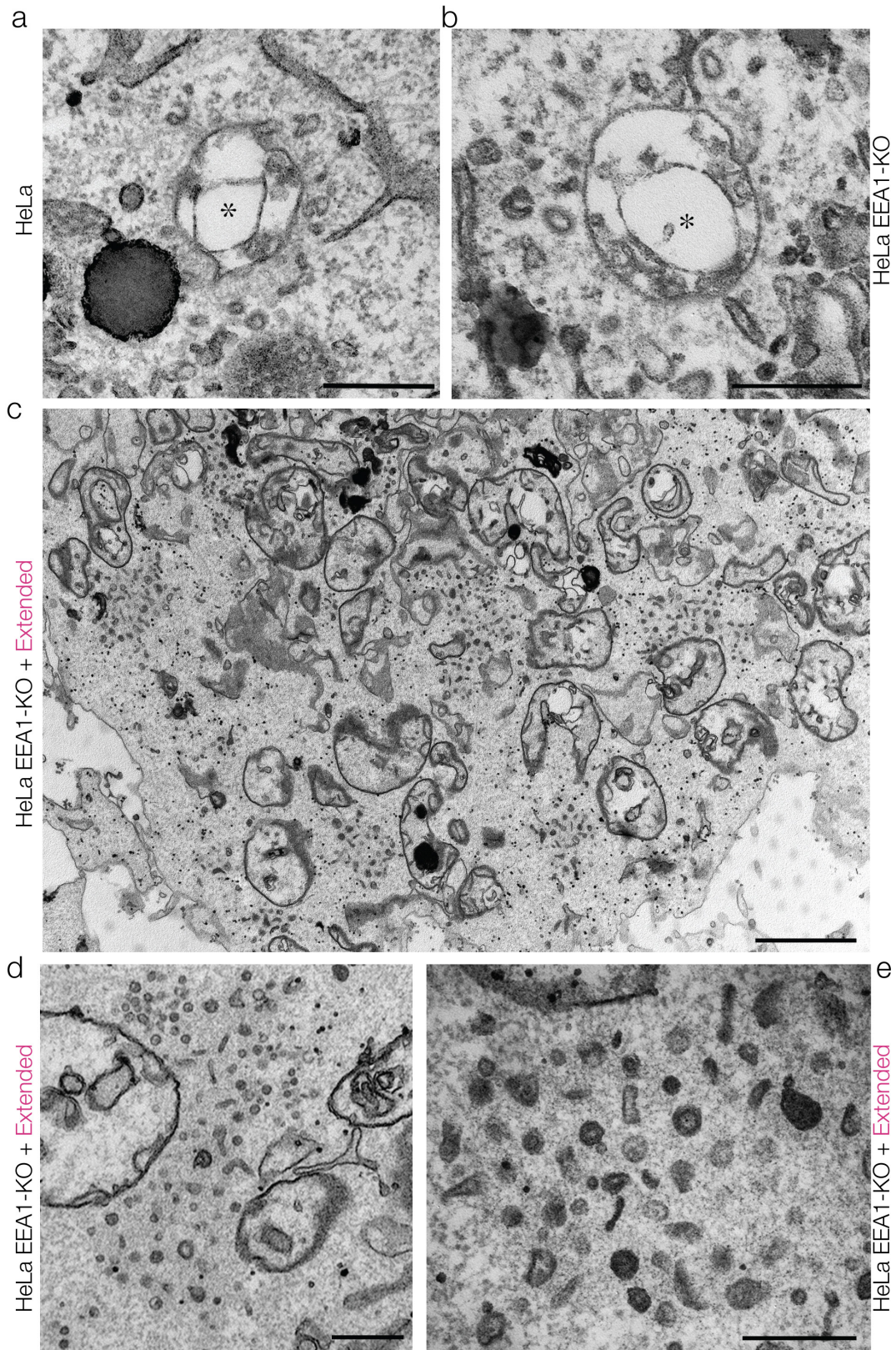


Extended Data Figure 8 | See next page for caption.

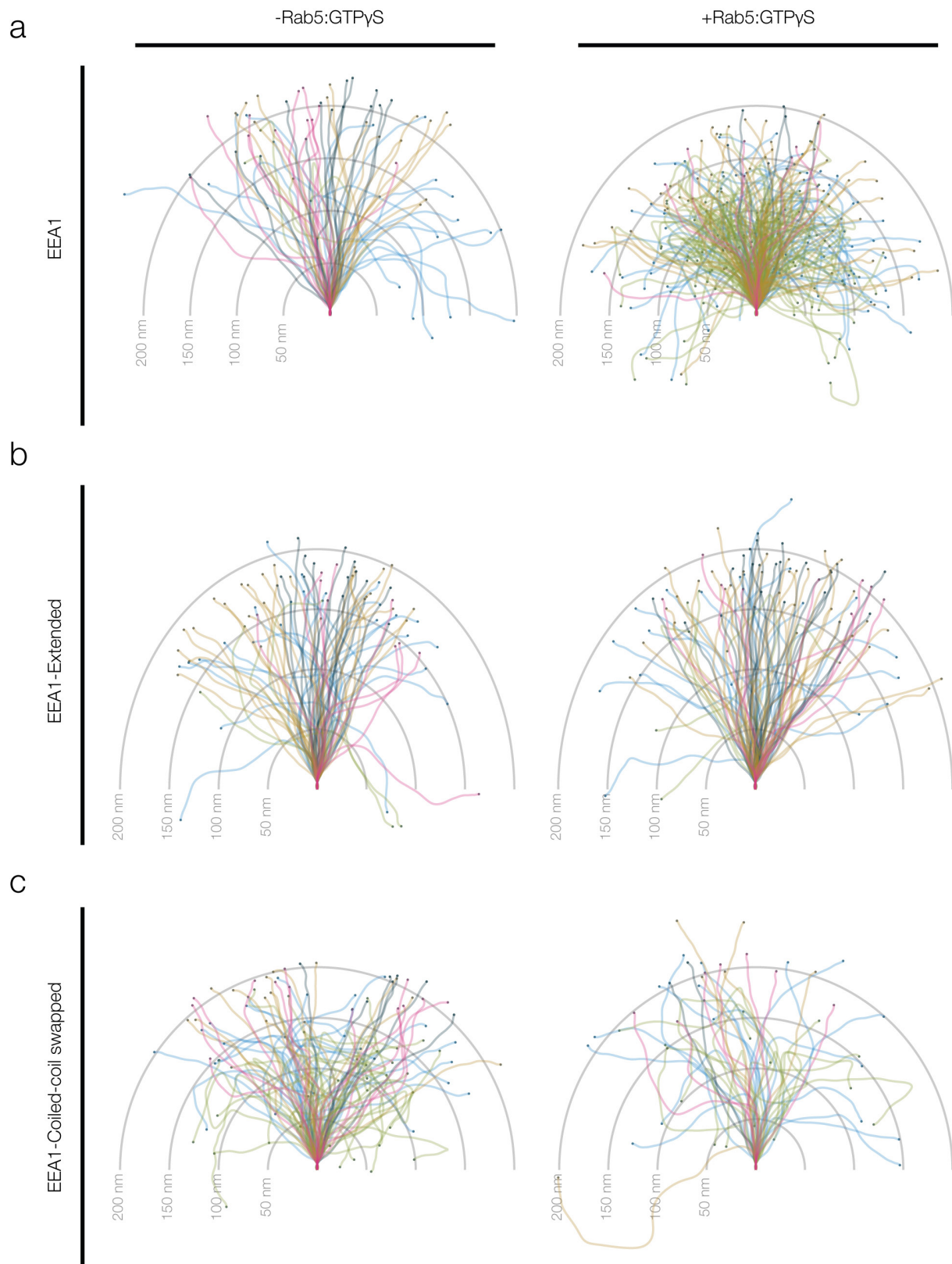
Extended Data Figure 8 | EEA1 mutants incapable of undergoing entropic collapse result in defects in endosomal trafficking.

a, b, Automated confocal immunofluorescence images ($n = 30$ each) of HeLa EEA1-KO and standard HeLa cells. EEA1 (green) and Rab5 (magenta). Scale bar, $10\ \mu\text{m}$. **c**, Western blot of HeLa and HeLa EEA1-KO clonal cell line for EEA1 and Rab5. **d, e, g, h**, Automated confocal images ($n = 30$ each) of HeLa EEA1-KO cells expressing no EEA1 (KO, **d**), rescued with wild-type EEA1 (rescue, **e**) or extended and swapped mutants (**g, h**). Cells were pulsed with fluorescently labelled cargo (LDL) (green) for 10 min, fixed and immunostained for Rab5 (magenta) and EEA1 (for EEA1, see Fig. 4). Magnified insets of endosomes are depicted at arrows. Scale bar, $10\ \mu\text{m}$. **f**, Relative complexity of Rab5 endosomes per cell. Each Rab5 endosome is segmented, and the segmented object requires a defined number of 2D Gaussian functions, hereby referred to as complexity. Relative to wild type, HeLa EEA1-KOs (black line) had a significantly reduced number of endosomes of high complexity (>3.0), but more endosomes defined simply by one or two Gaussian functions. Rescue experiments (red) revealed no significant difference in complexity. In contrast, both extended and swapped mutants (blue and green respectively) had significantly fewer simple endosomes of low complexity, and significantly more of higher complexity. Mean \pm s.d., $n = 30$.

i, Histogram of fluorescence intensity of EEA1 per cell. KO cell lines had a sharp peak of intensity at background levels, whereas wild-type HeLa cells had a normal distribution. Grey box represents threshold levels of EEA1 intensity per cell taken for analysis. **j–l**, EGF uptake experiments. Confocal images of HeLa EEA1-KOs expressing wild-type EEA1 (rescue, **j**) or extended and swapped mutants (**g, h**). Cells were pulsed with fluorescently labelled EGF (green) for 10 min, fixed and immunostained for EEA1 (magenta). Images shown are maximum intensity projections. Scale bar, $5\ \mu\text{m}$. **m**, HeLa EEA1-KO cells in which the swapped EEA1 mutant was reintroduced showed clusters of vesicles and more rarely the classical endosomal morphology. The clusters were clearly delineated by a zone of cytoplasm with a distinct density. Representative of $n = 19$. Scale bars, $2\ \mu\text{m}$. **n**, Further quantifications, and the swapped mutant ultrastructural phenotype. Fraction of endosomal surface containing filamentous material for HeLa and HeLa EEA1-KOs. Box-whisker plot with minimum/maximum values, $n = 22, 24$ endosomes. $**P < 0.01$, two-tailed Student's *t*-test. **o**, Distance measured between endosome and tethered vesicles (HeLa) or between vesicles within large clusters (extended) (surface-to-surface, $n = 158$ and 623 for HeLa and extended respectively; $***P < 10^{-4}$, two-tailed Student's *t*-test).



Extended Data Figure 9 | Unlabelled version of Fig. 5.



Extended Data Figure 10 | Bouquet plots of EEA1 and variants. EEA1 in the absence of Rab5 is predominantly extended. The initial five segments of the curves from rotary shadowing electron microscopy were aligned and the curves plotted with the end position highlighted (dots). Grey concentric hemispheres demarcate 50, 100, 150 and 200 nm extensions

from the origin. The end positions therefore resulted in a cloud of empirical positions for the EEA1 N terminus of EEA1 (left), and reveal the overall change in conformational space that can be occupied by EEA1 when bound to Rab5:GTP- γ S (right). **b**, Bouquet plots for the extended EEA1 variant. **c**, Bouquet plots for the swapped EEA1 variant.

An endosomal tether undergoes an entropic collapse to bring vesicles together

David H. Murray^{1*}, Marcus Jahnel^{1,2,3*}, Janelle Lauer¹, Mario J. Avellaneda^{1,2†}, Nicolas Brouilly¹, Alice Cezanne¹, Hernán Morales-Navarrete¹, Enrico D. Perini^{1,2}, Charles Ferguson⁴, Andrei N. Lupas⁵, Yannis Kalaidzidis¹, Robert G. Parton^{4,6}, Stephan W. Grill^{1,2,3} & Marino Zerial¹

An early step in intracellular transport is the selective recognition of a vesicle by its appropriate target membrane, a process regulated by Rab GTPases via the recruitment of tethering effectors^{1–4}. Membrane tethering confers higher selectivity and efficiency to membrane fusion than the pairing of SNAREs (soluble N-ethylmaleimide-sensitive factor attachment protein receptors) alone^{5–7}. Here we address the mechanism whereby a tethered vesicle comes closer towards its target membrane for fusion by reconstituting an endosomal asymmetric tethering machinery consisting of the dimeric coiled-coil protein EEA1 (refs 6, 7) recruited to phosphatidylinositol 3-phosphate membranes and binding vesicles harbouring Rab5. Surprisingly, structural analysis reveals that Rab5:GTP induces an allosteric conformational change in EEA1, from extended to flexible and collapsed. Through dynamic analysis by optical tweezers, we confirm that EEA1 captures a vesicle at a distance corresponding to its extended conformation, and directly measure its flexibility and the forces induced during the tethering reaction. Expression of engineered EEA1 variants defective in the conformational change induce prominent clusters of tethered vesicles *in vivo*. Our results suggest a new mechanism in which Rab5 induces a change in flexibility of EEA1, generating an entropic collapse force that pulls the captured vesicle towards the target membrane to initiate docking and fusion.

EEA1, as nearly all putative coiled-coil tethering proteins, extends more than ten times the length of SNARE proteins^{8,9}. To explain how such a long molecule can mediate membrane tethering but also allow the membranes to come closer for fusion, we reconstituted a minimal asymmetric membrane tethering in liposomes containing EEA1, Rab5 and different fluorescent tracers (Fig. 1a and Extended Data Fig. 1b–e). EEA1 binds to phosphatidylinositol 3-phosphate (PI(3)P) via its carboxy (C) terminus with high affinity (dissociation constant $K_d \approx 50$ nM)^{7,10–12}, and to Rab5:GTP via its amino (N) terminus with comparatively lower affinity ($K_d \approx 2.4$ μ M)¹³. Liposomes containing PI(3)P and labelled with RhoDPPE effectively recruited EEA1 and tethered to DiD-labelled Rab5-6 \times His-liposomes, as analysed by confocal microscopy (Fig. 1a–c). The reaction required EEA1, Rab5 and GTP- γ S, as no co-localization was observed in the presence of GDP. The efficiency of tethering approached that of biotin–streptavidin liposomes (Fig. 1d). Furthermore, no co-localization was observed between pairs of liposomes harbouring Rab5 (Fig. 1e). Therefore, Rab5, EEA1 and PI(3)P form a minimal endosomal asymmetric membrane tethering machinery.

In principle, the N terminus of EEA1 could also bind Rab5 *in cis*: that is, on the same membrane. However, the presence of Rab5 on both pairs of liposomes, as in early endosomes *in vivo*, did not interfere with the tethering activity of EEA1 *in vitro*, as tethering was

indistinguishable between the asymmetric and symmetric conditions (Fig. 1c, e). Moreover, coiled-coil prediction algorithms estimate a central segment of nearly ~ 200 nm (refs 14, 15) (Extended Data Fig. 1a), suggesting that the molecule adopts an extended conformation. Indeed, filamentous EEA1-positive structures emanating from the surface of early endosomes *in vivo* have been observed by electron microscopy¹¹. In further support of this interpretation, we visualized the N and C termini of EEA1 using specific antibodies by super-resolution microscopy in HeLa cells (Fig. 1f, g, Extended Data Fig. 1f–h and Methods). If the N terminus of EEA1 bound Rab5 *in cis*, it should co-localize with the C terminus. Strikingly, the ends of EEA1 could instead be resolved, with the N terminus extending radially from the C terminus into the cytoplasm. We estimated an end-to-end distance of 141 ± 47 nm (mean \pm s.d.; Fig. 1h), in the range of the predicted length and rigidity of coiled-coils.

To characterize the distances and dynamics of the tethering reaction, we generated bead-supported membranes (10 μ m silica microspheres) harbouring green fluorescent protein (GFP)–Rab5 (Fig. 1i and Extended Data Fig. 2). These tethered to liposomes containing PI(3)P in the presence of GTP- γ S but not GDP in an EEA1 concentration-dependent manner (Extended Data Fig. 2g, h). Time-lapse microscopy showed that some liposomes were captured by the bead-supported membrane, while others diffused away (Extended Data Fig. 2i and Supplementary Videos 1 and 2), similar to the behaviour of endosomes *in vivo*¹⁶. We next measured the distances between the tethered vesicle and GFP–Rab5 (Fig. 1j, Extended Data Fig. 2j and Methods). Surprisingly, we observed distances ranging from 20 nm up to approximately the predicted length of 200 nm (mean \pm s.d.; 84 ± 56 nm) (Fig. 1k). Such a broad distribution is irreconcilable with the predicted length of EEA1 and suggests that EEA1 may change its conformation.

We determined the conformation of EEA1 using rotary shadowing electron microscopy and image analysis (Fig. 2a). The measurements of contour length and mean end-to-end distance followed Gaussian distributions with an average of 222 ± 26 nm (Fig. 2b, top) and 195 ± 26 nm (Fig. 2b, bottom), respectively, confirming that the molecule is largely extended, as *in vivo*¹¹ (Fig. 1g, h). However, this is incompatible with the much shorter distances between tethered vesicles *in vitro* (Fig. 1k). Therefore, we asked whether binding to Rab5 may cause EEA1 to adopt a more compact conformation. Remarkably, this was the case. Addition of Rab5:GTP- γ S (Fig. 2c) resulted in a significant fraction of bent EEA1 molecules having a substantially reduced end-to-end distance of 122 ± 50 nm (Fig. 2d).

To gain further insights into this mechanism, we generated two mutants with alterations in the coiled-coil but retaining the Rab5- and PI(3)P-binding domains (Extended Data Fig. 3 and Methods). In the extended EEA1 mutant, we removed regions of discontinuity

¹Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstraße 108, 01307 Dresden, Germany. ²Biotechnology Center, Technical University Dresden, Tatzberg 47/49, 01307 Dresden, Germany. ³Max Planck Institute for the Physics of Complex Systems, Nöthnitzerstraße 38, 01187 Dresden, Germany. ⁴Institute for Molecular Bioscience, The University of Queensland, St Lucia 4072, Australia. ⁵Department of Protein Evolution, Max-Planck-Institute for Developmental Biology, 72076 Tübingen, Germany. ⁶Centre for Microscopy and Microanalysis, The University of Queensland, St Lucia 4072, Australia. [†]Present address: FOM Institute AMOLF, Science Park 104, 1098 XG Amsterdam, the Netherlands.

*These authors contributed equally to this work.

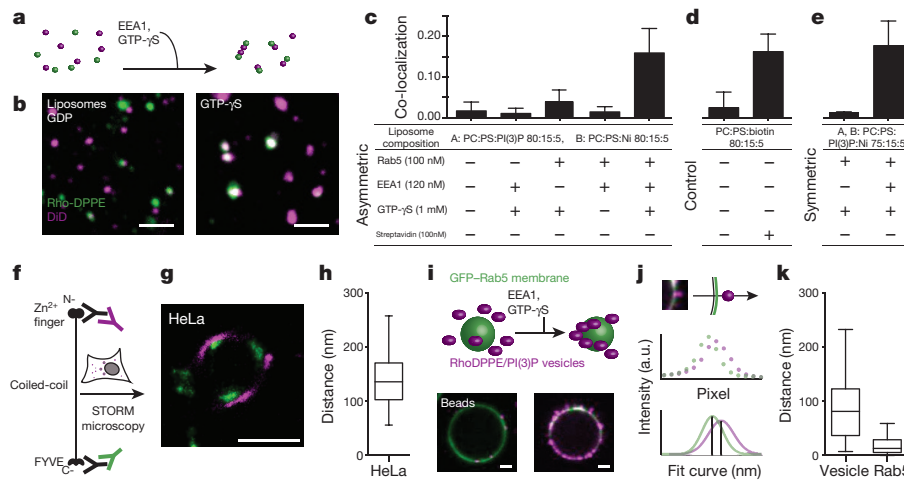


Figure 1 | EEA1, Rab5 and PI(3)P form an asymmetric tethering machinery. **a, b**, Vesicle–vesicle tethering assay. Rho–DPPE liposomes harbouring Rab5 (green) tether to DiD–PI(3)P liposomes (magenta) upon addition of EEA1 and GTP- γ S but not GDP (**a**, scheme; **b**, microscopy; representative of $n = 20$). Scale bar, $2\ \mu\text{m}$. **c–e**, Analysis of vesicle co-localization. Asymmetric (**c**) and symmetric (**e**) tethering required Rab5, PI(3)P and EEA1, streptavidin–biotin control (**d**) (mean \pm s.d., $n = 3$). **f–h**, *In vivo* stochastic optical reconstruction microscopy (STORM) defines the extension of EEA1. The N-terminal (magenta) and C-terminal (green) domains of EEA1 (**f**) were differentially labelled. Representative

STORM image (**g**, of $n = 22$) and quantification of EEA1 extension (**h**, box–whisker plot with median, 25/75 quartiles and minimum/maximum error bars, $n = 86$, representative experiment) from endosomes. Scale bar, $500\ \text{nm}$. **i**, Bead-supported membrane tethering similar to **a** and **b**. Representative of $n = 20$. Scale bar, $2\ \mu\text{m}$. **j, k**, Distance of tethered vesicles (magenta) from the membrane (green). The intensity per pixel was plotted, fitted to determine the relative distances and quantified (**k**) (vesicle–membrane and Rab5–membrane, representative experiment; box–whisker plot as in **h**, mean \pm s.d., $n = 36$ and 14).

between heptad repeats creating a more idealized, extended coiled-coil. In the swapped EEA1 mutant, we swapped the coiled-coil regions between the N and C termini. Electron microscopy analysis revealed that the extended mutant was impaired in the Rab5-induced conformational change (Fig. 2i and Extended Data Fig. 4a–c). In contrast, the swapped mutant was mostly bent, often presented kinks, and did not significantly change conformation upon Rab5 binding (Fig. 2f and Extended Data Fig. 4e–g). These results suggest that coiled-coil discontinuities and their physical arrangement are

critical for the structure of EEA1 and its Rab5-induced conformational change.

To shed light on how EEA1 adopts a compact conformation upon Rab5 binding, we measured the curvature along the contour of molecules. We aligned N-terminally MBP-tagged EEA1 and determined how the tangents to the contour change by $8\ \text{nm}$ steps along the contour (Methods and Extended Data Fig. 5). Interestingly, the variance of this measure of curvature calculated over the ensemble of molecules increased significantly upon Rab5:GTP- γ S binding (Fig. 2g), indicating

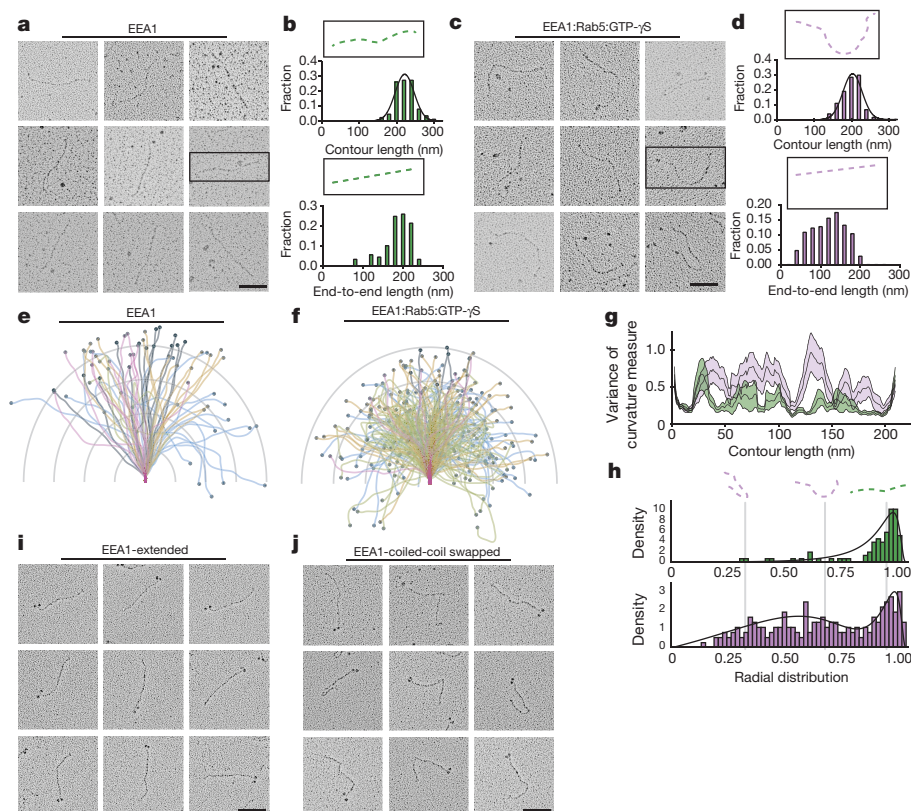


Figure 2 | EEA1 changes flexibility upon Rab5 binding. **a, c, i, j**, Representative examples of rotary-shadowing electron microscopy of EEA1 (**a**), EEA1 + Rab5:GTP- γ S (**c**), EEA1-extended (**i**) and -swapped (**j**) variants. Scale bar, $100\ \text{nm}$; $n = 88$, $n = 212$, $n = 90$, $n = 145$, respectively. **b, d**, Contour and end-to-end length histograms for EEA1 (green, $n = 88$) and EEA1 + Rab5:GTP- γ S (magenta, $n = 212$). **e, f**, Visual comparison of aligned EEA1 proteins. The highlighted ends of EEA1 + Rab5:GTP- γ S lie significantly closer to the origin. Hemispheres demarcate $50\ \text{nm}$. **g**, Variance of curvature measures along the contour of aligned EEA1 + Rab5:GDP (green) and EEA1 + Rab5:GTP- γ S (magenta) molecules ($n = 90$, $n = 145$, respectively). **h**, Radial distribution functions define the extension probability for EEA1 \pm Rab5:GTP- γ S (–Rab5:GTP- γ S, green; +Rab5:GTP- γ S, magenta) with fit (black lines).

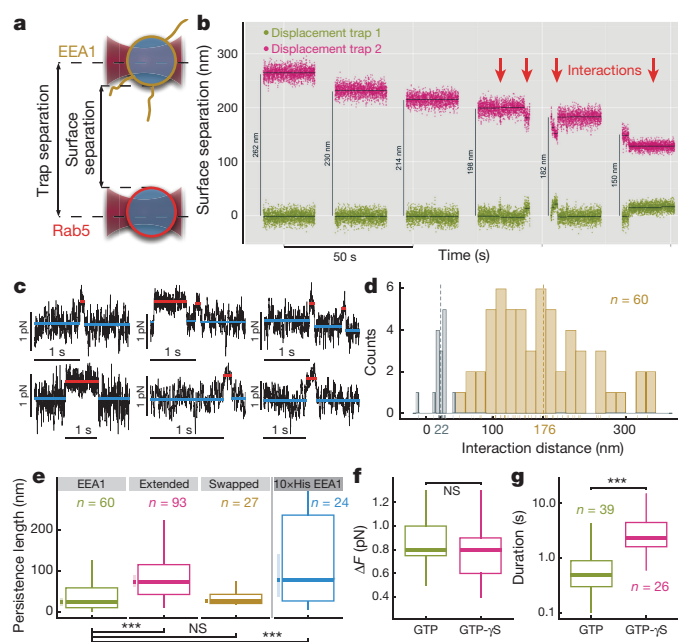


Figure 3 | EEA1 collapse generates a force. **a**, Scheme of bead-supported membranes harbouring EEA1 or Rab5 captured by dual-trap optical tweezers. **b**, **c**, Traps moved successively closer until interactions (arrows) were observed, characterized by increase in force and decrease in variance (c). **d**, Interaction distance consistent with length of extended EEA1. Silica microspheres (negative control) in grey. **e**, Persistence length distributions of EEA1 and variants from optical tweezers measurements. **f**, Force did not depend on GTP hydrolysis ($P > 0.15$); $n = 39, 26$ respectively. **g**, Interaction duration (log-scale) was prolonged by GTP- γ S ($P < 10^{-4}$). Mann–Whitney–Wilcoxon test (e–g); box–whisker plot with Tukey error bars (e–g).

that EEA1 displays a larger variety of curvatures upon Rab5:GTP binding. Such changes occurred along the entire length of the molecule, with some regions increasing in flexibility more than others (Fig. 2g), but were not observed for the EEA1 mutants (Extended Data Fig. 5f–i).

Although molecules are adsorbed onto a 2D surface, some aspects of their 3D conformations are captured (Methods). Analysis of the kurtosis of the distribution of angles between contour tangents indicated that 3D shape fluctuations are retained for the entire contour of EEA1 in the presence of Rab5:GDP, but only up to 60 nm with Rab5:GTP- γ S (Methods and Extended Data Fig. 6). Moreover, tangent–tangent correlations of the contour in this regime revealed that Rab5:GTP- γ S binding results in a faster decay. Generally, the worm-like chain (WLC) model is used to describe fluctuations in polymer shapes and

capture aspects of the physics underlying their shape fluctuations¹⁷ (Methods). In the WLC model, the polymer is considered a homogeneous molecule with its flexibility determined by a bending stiffness reflected in a characteristic length, the persistence length, over which correlations between tangents to the contour decay. We applied the WLC model to EEA1 and determined an effective persistence length of 246 ± 42 nm for the unbound and 74 ± 3 nm for the Rab5:GTP- γ S-bound ensembles. In contrast, the extended EEA1 mutant had similar effective persistence lengths in either state (unbound = 183 ± 13 nm and bound = 224 ± 25 nm; Supplementary Data Table).

To corroborate these estimates, we fitted the radial distribution functions (that is, the probability of observing a given end-to-end distance) of the molecules extracted from the electron microscopy data with analytical solutions of the WLC model¹⁸ (Methods). This showed a clear reduction in effective persistence length of EEA1 upon Rab5:GTP binding (Fig. 2h). In contrast, the extended EEA1 mutant maintained a similar radial distribution regardless of Rab5 (Extended Data Fig. 4d).

Reducing the persistence length of EEA1 makes the molecule flexible. However, the tether is still extended and, therefore, in an out-of-equilibrium conformation (Fig. 2e). As a result, it will undergo an entropic collapse, with its end-to-end distance decreasing towards a new equilibrium (Fig. 2f). This process generates a force that could pull the membranes together (estimated ~ 3 pN (Methods)). In some sense, the extended molecule is like a loaded spring that rapidly recoils upon Rab5 binding.

To provide experimental evidence for entropic collapse of EEA1, we made use of high-resolution dual-trap optical tweezers (Methods). Two glass $2\mu\text{m}$ microspheres coated with membranes were held in optical traps (Fig. 3a). One trap was moved closer to the other, in iterative cycles of approaching, pausing and retracting (Fig. 3b). At distances below 250 nm and at low concentrations of EEA1 (5–40 nM) to ensure single-molecule events, we observed transient interactions as a decrease in the mean and variance of the distance between the two beads (Fig. 3b, red arrows, Fig. 3c and Extended Data Fig. 7a, d). Interactions were infrequent, as expected for single molecules and non-existent without EEA1, whereas their frequency and duration increased at high concentrations of EEA1 (400 nM) (Extended Data Fig. 7e and Methods). The interaction distance was broad (Fig. 3d), with the mean 176 ± 76 nm comparing favourably with rigid EEA1 (Fig. 2b).

To test the prediction that EEA1 becomes flexible upon Rab5 binding, for each tethered molecule we determined its effective persistence length from the capture distance, and measured force increase (Fig. 3c) and bead displacements using the WLC model (Methods). Strikingly, we obtained a median effective persistence length of 23 ± 10 nm (Fig. 3e). For more than 80% of the molecules the persistence length was no more than half of the contour length, confirming that Rab5-bound

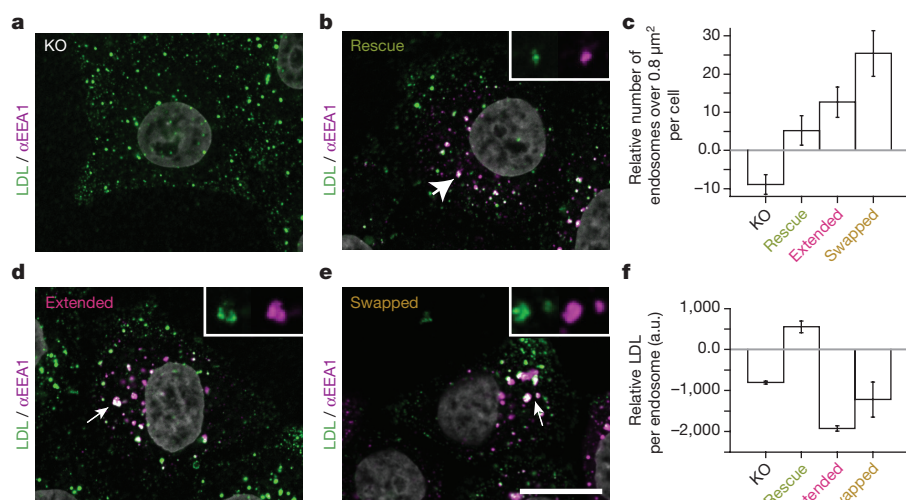


Figure 4 | EEA1 mutants blocking entropic collapse induce trafficking defects. **a**, **b**, **d**, **e**, Confocal images of HeLa EEA1-KO cells (a), rescued with EEA1, extended or swapped mutants (b, d, e). Uptake of LDL (green) and immunostaining for EEA1 (magenta). Inset, endosomes depicted at arrows. Representative of $n = 30$ images per condition (Methods). Scale bar, $10\mu\text{m}$. **c**, **f**, Relative difference in number of large endosomes (c) and LDL fluorescence (f) (a.u., arbitrary units). Mean \pm s.d., representative experiment of 3, $n = 30$ images. $P < 0.01$ versus HeLa, t -test, except rescue.

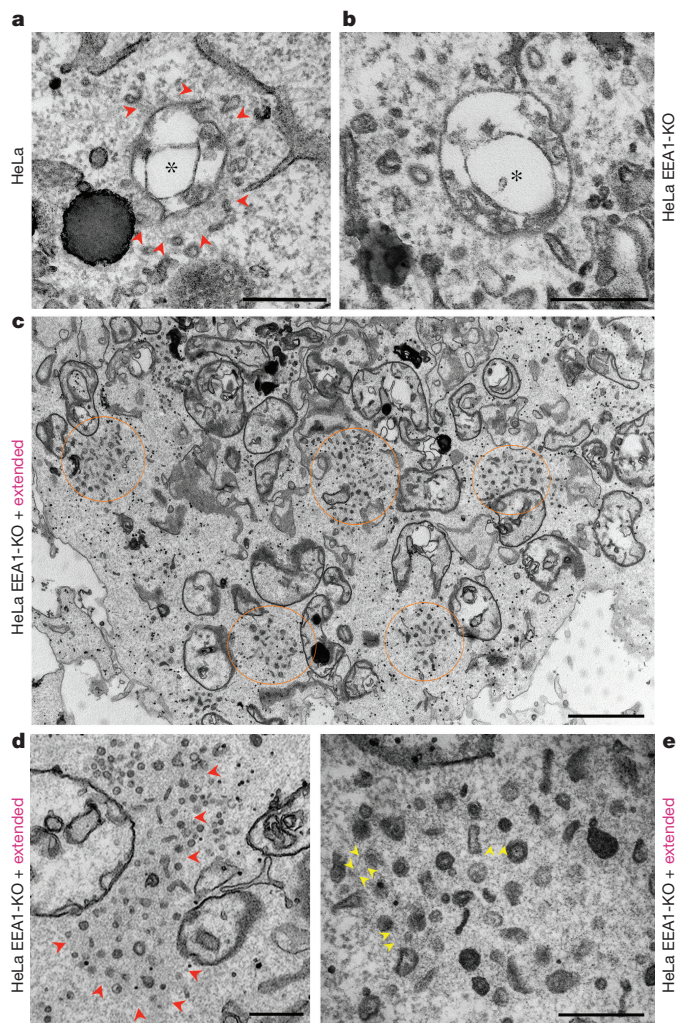


Figure 5 | Ultrastructural analysis of EEA1 KO and mutant rescue cells. **a**, Dense filamentous network (arrowheads) around an early endosome (asterisks) in HeLa. Many smaller vesicular or tubular profiles were consistently observed at the network periphery. Representative of $n = 33$. **b**, A filamentous network was less prominent in HeLa EEA1-KO with no obvious concentration of vesicles near the endosomal surface. Representative of $n = 54$. **c–e**, HeLa EEA1-KO expressing the extended EEA1 variant showed clusters of vesicles throughout the cytoplasm and no classical endosomal morphology. The clusters were clearly delineated by a zone of cytoplasm with distinct density (circled areas). Higher magnification revealed fine wispy material surrounding the clustered vesicles (**d**, **e**; arrowheads) and evidence of discrete filaments (between the arrowheads in **e**). Representative of $n = 56$. Scale bars: **a**, **b**, **d**, **e**, 500 nm; **c**, 2 μm .

EEA1 is flexible. In contrast, the extended EEA1 mutant remained significantly more rigid than EEA1 (Fig. 3e). Rab5 binding is necessary to trigger structural and conformational changes on EEA1. When Rab5 was bypassed by His-tag-mediated tethering, EEA1 flexibility was significantly lower than that of EEA1 with Rab5 (Fig. 3e).

If EEA1 becomes flexible upon capture, an entropic pulling force will be generated. This entropic force balances with the force exerted by the optical traps as the molecule undergoes the collapse and as the system finds its new equilibrium (Extended Data Fig. 7h)¹⁹. For a capture distance of 195 nm and a peak collapse force of 3 pN, we predict a force balance at ~ 0.6 pN (Methods), consistent with our tweezer measurements of 0.5 ± 0.3 pN (Fig. 3c). EEA1 binding to Rab5 requires the GTP-bound form. No significant force differences were observed in the presence of the non-hydrolysable analogue GTP- γ S or GTP (Fig. 3f). In contrast, the duration of the interaction was much

prolonged (Fig. 3g), as expected given that GTP- γ S stabilizes Rab5 in the active form²⁰. Finally, replacing EEA1-Rab5 binding with $10\times$ His-EEA1 tethering to Ni-NTA-beads resulted in a decreased collapse force (Extended Data Fig. 7i).

To validate *in vivo* the mechanism observed *in vitro*, we genome-edited HeLa cells to disrupt the EEA1 gene (HeLa EEA1-KO; Fig. 4a, Extended Data Fig. 8c and Methods), and analysed the distribution of Rab5-positive endosomes and the uptake of cargo (low-density lipoprotein (LDL)) by confocal microscopy (Fig. 4a). HeLa EEA1-KO displayed a significant reduction in Rab5 endosome size, particularly for the largest endosomes (Fig. 4c), and a marked decrease in cargo (LDL) uptake (Fig. 4f). Expression of EEA1 rescued the normal, rounded morphology of endosomes (Fig. 4b and Extended Data Fig. 8f, i) and LDL uptake (Fig. 4c). In contrast, the expression of both extended and swapped EEA1 mutants generated enlarged endosomes and inhibited cargo uptake (Fig. 4c–f).

Because the size of endosomes is below the resolution limit of light microscopy, we performed electron microscopy on the HeLa EEA1-KO cells (Fig. 5 and Extended Data Fig. 9). The filamentous material on endosomes¹¹ was much reduced in HeLa EEA1-KO cells (Fig. 5a, b, and Extended Data Fig. 8n) and restored by the re-expression of EEA1 on endosomes that appeared normal or enlarged, consistent with the light microscopy analysis (Fig. 4b). Strikingly, cells expressing the extended EEA1 mutant had large ($>1\mu\text{m}$) clusters of small vesicles, within areas filled with filamentous material (Fig. 5d, e), suggesting that they are arrested in a tethered state (Fig. 4d, e). The distance between the tethered vesicles was significantly longer than that between endosomes in control cells (Extended Data Fig. 8o), consistent with the mutant EEA1 being incapable of undergoing entropic collapse to shorter distances (Figs 2e and 3e). Similar endosomal clusters were induced by the swapped mutant (Extended Data Fig. 8m).

Our data suggest a new mechanochemical cycle of EEA1 regulated by Rab5:GTP binding and GTP hydrolysis. On early endosomes, EEA1 is in the extended state (Fig. 2e) and increases the probability of capturing a vesicle bearing Rab5. Similarly, it forms a Rab5-selectivity barrier (analogous to a polymer brush)²¹. When Rab5 on an incoming vesicle binds EEA1, it induces an allosteric conformational change, from extended to flexible (Fig. 2f). This shows a new function of Rab proteins beyond effector recruitment. The reduction in persistence length of EEA1 causes its entropic collapse, releasing up to $\sim 14 k_B T$ of mechanical energy (Extended Data Fig. 7k) and generating up to 3 pN of force that could pull the vesicle closer to its target membrane where it may diffuse²² or be brought by other Rab5 effectors^{23,24} within the range of trans-SNARE pairing. This mechanism explains why the Rab5 machinery dramatically increases the efficiency of SNARE-mediated membrane fusion²³. The mechanical energy released by EEA1 is of the order of the free energy released by GTP hydrolysis. However, the energy required to complete the cycle could potentially also come from chaperones.

A key question is how Rab5 can induce such a long-range allosteric effect. This is not uncommon among coiled-coil proteins^{25,26}. The entropic collapse mechanism is different, however, for other membrane tethering factors²⁷. In the course of this study, the GCC185 tether was shown to bend through central joints²⁷. For EEA1, instead (1) the arrangement and structure of the coiled-coils and (2) Rab5 binding are critical for the propagation of allosteric conformational changes (Extended Data Fig. 10). We can envisage different mechanisms (see Supplementary Discussion), such as local register shifts. In dynein, dynamics in the heptad register prove critical to functionally link ATP binding and microtubule binding at opposite ends of its coiled-coil stalk^{28,29}. Further ad hoc structural studies are necessary to resolve this outstanding problem. The entropic collapse upon stiffness reduction could be an effective and general mechanism used not only by membrane tethers but also by many coiled-coil proteins for generating an attractive force in diverse biological processes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 September 2015; accepted 19 July 2016.

Published online 24 August 2016.

1. Bröcker, C., Engelbrecht-Vandré, S. & Ungermann, C. Multisubunit tethering complexes and their role in membrane fusion. *Curr. Biol.* **20**, R943–R952 (2010).
2. Brown, F. C. & Pfeffer, S. R. An update on transport vesicle tethering. *Mol. Membr. Biol.* **27**, 457–461 (2010).
3. Zerial, M. & McBride, H. Rab proteins as membrane organizers. *Nature Rev. Mol. Cell Biol.* **2**, 107–117 (2001).
4. Munro, S. Organelle identity and the organization of membrane traffic. *Nature Cell Biol.* **6**, 469–472 (2004).
5. Mayer, A. & Wickner, W. Docking of yeast vacuoles is catalyzed by the Ras-like GTPase Ypt7p after symmetric priming by Sec18p (NSF). *J. Cell Biol.* **136**, 307–317 (1997).
6. Christoforidis, S., McBride, H. M., Burgoyne, R. D. & Zerial, M. The Rab5 effector EEA1 is a core component of endosome docking. *Nature* **397**, 621–625 (1999).
7. Rubino, M., Miaczynska, M., Lippé, R. & Zerial, M. Selective membrane recruitment of EEA1 suggests a role in directional transport of clathrin-coated vesicles to early endosomes. *J. Biol. Chem.* **275**, 3745–3748 (2000).
8. Gao, Y. *et al.* Single reconstituted neuronal SNARE complexes zipper in three distinct stages. *Science* **337**, 1340–1343 (2012).
9. Kiessling, V. & Tamm, L. K. Measuring distances in supported bilayers by fluorescence interference-contrast microscopy: polymer supports and SNARE proteins. *Biophys. J.* **84**, 408–418 (2003).
10. Dumas, J. J. *et al.* Multivalent endosome targeting by homodimeric EEA1. *Mol. Cell* **8**, 947–958 (2001).
11. Wilson, J. M. *et al.* EEA1, a tethering protein of the early sorting endosome, shows a polarized distribution in hippocampal neurons, epithelial cells, and fibroblasts. *Mol. Biol. Cell* **11**, 2657–2671 (2000).
12. Simonsen, A. *et al.* EEA1 links PI(3)K function to Rab5 regulation of endosome fusion. *Nature* **394**, 494–498 (1998).
13. Mishra, A., Eathiraj, S., Corvera, S. & Lambright, D. G. Structural basis for Rab GTPase recognition and endosome tethering by the C2H2 zinc finger of early endosomal autoantigen 1 (EEA1). *Proc. Natl Acad. Sci. USA* **107**, 10866–10871 (2010).
14. Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164 (1991).
15. McDonnell, A. V., Jiang, T., Keating, A. E. & Berger, B. Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics* **22**, 356–358 (2006).
16. Rink, J., Ghigo, E., Kalaidzidis, Y. & Zerial, M. Rab conversion as a mechanism of progression from early to late endosomes. *Cell* **122**, 735–749 (2005).
17. Landau, L. D. & Lifshitz, E. M. *Statistical Physics* 3rd edn, Part 1, Vol. 5, Ch. 12, 396–400 (Butterworth-Heinemann, 1980).
18. Wilhelm, J. & Frey, E. Radial distribution function of semiflexible polymers. *Phys. Rev. Lett.* **77**, 2581–2584 (1996).
19. Otto, O., Sturm, S., Laohakunakorn, N., Keyser, U. F. & Kroy, K. Rapid internal contraction boosts DNA friction. *Nature Commun.* **4**, 1780 (2013).
20. Rybin, V. *et al.* GTPase activity of Rab5 acts as a timer for endocytic membrane fusion. *Nature* **383**, 266–269 (1996).
21. Milner, S. T. Polymer brushes. *Science* **251**, 905–914 (1991).
22. Degtyar, V. E., Allersma, M. W., Axelrod, D. & Holz, R. W. Increased motion and travel, rather than stable docking, characterize the last moments before secretory granule fusion. *Proc. Natl Acad. Sci. USA* **104**, 15929–15934 (2007).
23. Ohya, T. *et al.* Reconstitution of Rab- and SNARE-dependent membrane fusion by synthetic endosomes. *Nature* **459**, 1091–1097 (2009).
24. Perini, E. D., Schaefer, R., Stöter, M., Kalaidzidis, Y. & Zerial, M. Mammalian CORVET is required for fusion and conversion of distinct early endosome subpopulations. *Traffic* **15**, 1366–1389 (2014).
25. Moreno-Herrero, F. *et al.* Mesoscale conformational changes in the DNA-repair complex Rad50/Mre11/Nbs1 upon binding DNA. *Nature* **437**, 440–443 (2005).
26. Taylor, K. C. *et al.* Skip residues modulate the structural properties of the myosin rod and guide thick filament assembly. *Proc. Natl Acad. Sci. USA* **112**, E3806–E3815 (2015).
27. Cheung, P. Y., Limouse, C., Mabuchi, H. & Pfeffer, S. R. Protein flexibility is required for vesicle tethering at the Golgi. *eLife* **4**, e12790 (2015).
28. Schmidt, H., Zalyte, R., Urnavicius, L. & Carter, A. P. Structure of human cytoplasmic dynein-2 primed for its power stroke. *Nature* **518**, 435–438 (2015).
29. Kon, T. *et al.* Helix sliding in the stalk coiled coil of dynein couples ATPase and microtubule binding. *Nature Struct. Mol. Biol.* **16**, 325–333 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank R. Schäfer for project support. We acknowledge discussions with J. Morin, Ü. Coskun, A. Honigsmann, S. Sturm and T. Leonard, and F. Jülicher, E. Schäfer and K. Simons for reading the manuscript. We thank M. Brandstetter and the electron microscopy facility of the Vienna Biocenter. We thank the Light Microscopy, Protein Expression, Chromatography, and High-throughput Technology Development Studio of the Max Planck Institute of Molecular Cell Biology and Genetics. During part of the work, M.J. was supported by a PhD scholarship of the Böhringer Ingelheim Fonds. M.J.A. was supported by the La Caixa and Deutscher Akademischer Austauschdienst scholarship. R.P. was supported by the National Health and Medical Research Council of Australia (program grant APP1037320 and Senior Principal Research Fellowship 569452) and the Australian Research Council Centre of Excellence (CE140100036). We acknowledge the Australian Microscopy & Microanalysis Research Facility at the Center for Microscopy and Microanalysis at The University of Queensland. S.W.G. was supported by the Deutsche Forschungsgemeinschaft (SPP 1782, GSC 97, GR 3271/2, GR 3271/3, GR 3271/4), the European Research Council (grant number 281903) and the Human Frontier Science Program (RGP0023/2014). This research was supported by the Max Planck Society and funds of the Deutsche Forschungsgemeinschaft (Transregio 83).

Author Contributions D.H.M., M.J., S.W.G. and M.Z. conceived the project together. D.H.M. prepared all reagents, performed experiments and their analysis. M.J. and S.W.G. interpreted data in the context of polymer physics. M.J. performed optical tweezer experiments with D.H.M. M.J.A. and E.P. performed initial tweezer experiments. M.J. and M.J.A. analysed tweezer experiments. D.H.M., J.L. and A.N.L. designed mutants. N.B. performed super-resolution experiments. A.C. assisted in reconstitution experiments. C.F. and R.G.P. performed cell electron microscopy, and D.H.M., H.M.-N. and M.J. analysed electron microscopy data. Y.K. analysed cell microscopy. D.H.M., M.J., S.W.G. and M.Z. wrote the manuscript with input from all the authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.W.G. (stephan.grill@biotec.tu-dresden.de) or M.Z. (Zerial@mpi-cbg.de).

Reviewer Information *Nature* thanks C. Schmidt, J. Zimmerberg and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Statistics. Sample size was not predetermined. For cell electron microscopy, samples were double-blind examined. Other experiments were not randomized or blinded. Box-whisker plots all show median, 25/75 quartiles by box boundaries and minimum/maximum values by errors, with the exception of Fig. 3 and Extended Data Fig. 7 which use Tukey-defined error bars.

Cloning, expression and purification of proteins. Human Rab5-6 \times His and GFP-Rab5-6 \times His were expressed and purified essentially as previously described in the *Escherichia coli* expression system⁶. Human Rabex-5 amino-acid residues 131–394 were PCR and restriction cloned into a pGST-parallel2 vector containing a TEV cleavable N-terminal glutathione-S-transferase (GST)^{29,30}. Expression and purification was performed essentially as described³¹. Briefly, *E. coli*-expressed proteins were transformed into BL21(DE3) cells and grown at 37 °C until absorbance at 600 nm ($A_{600\text{ nm}}$) of 0.8, whereupon the incubator was reduced to 18 °C. After 30 min, cultures were induced with 0.1 mM IPTG and grown overnight (16 h). Cell pellets were resuspended in standard buffer (20 mM Tris pH7.4, 150 mM NaCl, 0.5 mM TCEP) and flash frozen in liquid nitrogen. All subsequent steps performed at 4 °C or on ice. Cell pellets were resuspended in standard buffer supplemented with 1 mM MgCl₂ for GTPases, and protease inhibitor cocktail (chymostatin 6 μ g/ml, leupeptin 0.5 μ g/ml, antipain-HCl 10 μ g/ml, aprotinin 2 μ g/ml, pepstatin 0.7 μ g/ml, APMSF 10 μ g/ml), homogenized and lysed by sonication. Histidine-tagged proteins were bound in batch to Ni-NTA resin in the presence of 20 mM imidazole, and eluted with 200 mM imidazole. GST-tagged proteins were purified on GS resin (GS-4B, GE Healthcare) by binding for 2 h followed by stringent washing, and cleavage from resin overnight. Imidazole-containing samples were immediately diluted after elution and tags cleaved during overnight dialysis. Following dialysis and tag cleavage, samples were concentrated and TEV or HRV 3C protease was removed by reverse purification through Ni-NTA or GS resin. Samples were then purified by size-exclusion chromatography on Superdex 200 columns in standard buffer.

Human EEA1 was purified as a GST fusion in a pOEM series vector (Oxford Expression Technologies) modified to contain a HRV 3C-cleavable N-terminal GST and protease cleavage site or from a modified pFastbac1 vector (Thermo Fisher Scientific)²³. Some samples were also purified as 6 \times His-MBP and 10 \times His fusions from a modified pOEM vector (rotary shadowing for N-to-C terminus alignment, and optical tweezer control, respectively; all other experiments performed with tags removed). Mutants were purified identically to wild-type EEA1.

SF9 cells growing in ESF921 media (Expression Systems) were co-transfected with linearized viral genome and the expression plasmid and selected for high infectivity. P1 and P2 virus was generated according to the manufacturer's protocol, and expression screens and time courses performed to optimize expression yield. Best viruses were used to infect 1–2 l SF9 cells at 10⁶ cells/ml at 1% vol/vol and routinely harvested after 40–48 h at about 1.5 \times 10⁶ cells/ml, suspended in standard buffer and flash frozen in liquid nitrogen. Pellets were thawed on ice and lysed by Dounce homogenizer. Purification took place rapidly in standard buffer at 4 °C on GS resin in batch format. Bound protein was washed thoroughly and cleaved from resin by HRV 3C protease overnight. Proteins retaining 6 \times His-MBP tags were purified on amylose resin and eluted with 10 mM maltose. Protein retaining 10 \times His were eluted from Ni-NTA resin in standard buffer supplemented with 200 mM imidazole. All EEA1 and mutants were immediately further purified by Superose 6 size-exclusion chromatography where they eluted as a single peak. All experiments were performed with a preparation confirmed for Rab5 and PI(3)P binding. Concentrations were determined by UV280 and Bradford assay. All proteins were aliquoted and flash frozen in liquid nitrogen and stored at –80 °C.

EEA1 variants extended and swapped were synthesized genes optimized for insect cell expression (Genscript). The extended mutant has regions of low coiled-coil prediction removed, resulting in an EEA1 construct 1,286 amino acids in length (versus 1,411 in wild-type EEA1) (see Extended Data Fig. 3). The swapped mutant has the C-terminal portion of the coiled-coil rearranged to follow the N-terminal Zn²⁺-finger domains, and the N-terminal portion of the coiled-coil therefore rearranged to the C-terminal region of EEA1. Variants were treated identically to wild-type EEA1 in purification.

Static light scattering. An autosampler equipped Viskotek TDAMax system was used to analyse the light-scattering from purified EEA1. Sample was loaded the autosampler and passed through a TSKGel G5000PW column (Tosoh Biosciences) and fractions were subjected to scattering data acquisition. Data obtained were averaged across the protein elution volume and molecular masses determined in OmniSEC software package.

Lipids. The following lipids were purchased and used directly: DOPC, DOPS, DOGS-NiNTA, RhoDPPE (Avanti), DiD (Invitrogen) and PI(3)P (Echelon Biosciences). Lipids were dissolved in chloroform, except PI(3)P in 1:2:0.8 CHCl₃:MeOH:H₂O. All were stored at –80 °C.

Rab5/PI(3)P binding by EEA1. Early endosome fusion assay was performed as previously described³². To assess the ability of EEA1 to bind competently in a GTP-dependent manner to Rab5, Rab5 was bound to GS resin and subsequently loaded with nucleotide (GDP, GTP- γ S) as previously described⁶. Binding of EEA1 and all variants to immobilized Rab5 proceeded for 1 h at room temperature, and the washed Rab5 resin was evaluated for EEA1 binding by western blot. Similarly, the binding of EEA1 to PI(3)P containing liposomes was evaluated as previously described by formation of liposomes composed of DOPC:DOPS or DOPC:DOPS:PI(3)P (85:15 or 80:15:5 respectively)³³. Briefly, liposomes were formed from the hydration of lipids at 1 mM in standard buffer, and combined with EEA1 for 1 h before ultracentrifugation to separate supernatant and pellet for western blotting to evaluate EEA1 sedimentation. Rabbit anti-EEA1 antibody was made in our laboratory.

Preparation of liposomes. Liposomes were formed by extrusion as previously described³⁴. Liposome compositions for fluorescence microscopy tethering assays were DOPC:DOPS:DOGS-NiNTA, DOPC:DOPS:PI(3)P, DOPC:DOPS:biotin-DPPE, with RhoDPPE and DiD where applicable. Liposome compositions for bead-supported membranes were DOPC:DOPS:DOGS-NiNTA, DOPC:DOPS:PI(3)P. Solvent was evaporated under nitrogen and vacuum overnight. The resulting residue was suspended in standard buffer, rapidly vortexed, freeze-thawed five times by submersion in liquid N₂ followed by water at 40 °C, and extruded by 11 passes through two polycarbonate membranes with a pore diameter of 100 nm (Avestin). Vesicles stored at 4 °C were used within 5 days.

Bead-supported bilayer preparation. Silica beads (2 μ m NIST-traceable size-standards for optical tweezers, or 10 μ m standard microspheres for microscopy; Corpustar) were thoroughly cleaned in pure ethanol and Hellmanex (1% sol., Hellma Analytics) before storage in water. Supported bilayers were formed as previously described with modifications³⁵. Liposomes composed of DOPC:DOPS 85:15 (with 5% PI(3)P and DOGS-NiNTA where applicable) were added to a solution containing 250 mM NaCl for tethering assays (10 μ m) and 100 mM for optical tweezers (2 μ m), and 5 \times 10⁶ beads. Liposomes were added to final concentration of 100 μ M and incubated for 30 min (final volume 100 μ l). Samples were washed with 20 mM Tris pH7.4 three times by addition of 1 ml followed by gentle centrifugation (at 380g). Final wash was with standard buffer. Salt concentrations were optimized by examination of homogeneity at the transverse plane followed by examination of the excess membrane at the coverslip plane (see Extended Data Fig. 2a–d). We found that the membranes were extremely robust in conditions where the bilayer is fully formed, and could be readily pipetted and washed, consistent with previous reports³⁶. Membrane-coated beads were used within 1 h of production and always stored before use on a rotary suspension mixer.

Confocal microscopy of vesicle-vesicle tethering assay. Glass coverslips were cleaned in ethanol, Hellmanex and thoroughly rinsed in water. In these experiments, the following concentrations were used: 1 nM Rabex-5 (131–394), 100 nM Rab5-6 \times His, 120 nM EEA1. Experiments were performed in standard buffer with 5 mM MgCl₂ and 1 μ M nucleotide. Liposomes and proteins were pre-mixed in low-binding tubes at concentrations indicated, incubated for 5 min and imaged immediately upon addition to the coverslip. Images were acquired with a Nikon TiE equipped with a 60 \times plan-apochromat 1.2 numerical aperture W objective and Yokagawa CSU-X1 scan head. Images were acquired on an Andor DU-897 back-illuminated CCD. Acquired images were processed by the SQUASH package for Fiji³⁷.

Confocal microscopy of bead-supported membrane tethering assay. A 200 μ l observation chamber (μ -Slide 8 well, uncoated, #1.5, ibidi) was pre-blocked with BSA (1 mg/ml in standard buffer) for 1.5–2 h and washed thoroughly. Finally, 180 μ l of standard buffer containing beads was added to the sample chamber. In these experiments, the following concentrations were used: 1 nM Rabex-5 (131–394), 100 nM GFP-Rab5-6 \times His, and the given EEA1 concentrations (between 30 and 400 nM). Nucleotide control experiments were performed at 190 nM EEA1. Experiments were performed in standard buffer with 2 mM MgCl₂ and 1 mM nucleotide. Altogether Rab5, Rabex5, nucleotide, EEA1 and buffer were mixed in low-binding tubes at concentrations indicated, and were added to 240 μ l final volume to assure mixing throughout the chamber volume.

Images for co-localization analysis were acquired with a Nikon TiE equipped with a 60 \times plan-apochromat 1.2 numerical aperture W objective and Yokagawa CSU-X1 scan head. Images were acquired on an Andor DU-897 back-illuminated CCD. Acquired images were processed by the SQUASH package for Fiji³⁷.

Data obtained for distance measurements were acquired in the same way and processed in Fiji by determining line profiles eight pixels wide from the centre of the bead outwards over an observed vesicle. These profiles were fitted with a Gaussian distribution. The alignment of the microscope was confirmed by imaging of sub-diffraction beads, revealing no clear systematic shift and a maximum positional error of 21 nm determined in Motion Tracking¹⁶. Controls with sub-diffraction-sized multicolour particles (Methods) and distance measurements between Rab5

itself and its resident membrane were within the measurement error of the technique (approximately 15 nm)³⁸.

Super-resolution imaging of EEA1 termini. HeLa cells were stained using primary antibodies against EEA1 N terminus (610457, prepared in mouse, BD Biosciences) and EEA1 C terminus (2900, prepared in rabbit, Abcam). The secondary antibodies were anti-mouse Alexa568 antibody (A-11004, prepared in goat, Life Technologies) and anti-rabbit Alexa647 (A-21244, prepared in goat, Life Technologies). Coverslips were mounted in STORM buffer (100 mM Tris-HCl pH8.7, 10 mM NaCl, 10% glucose, 15% glycerol, 0.5 mg/ml glucose oxidase, 40 µg/ml catalase, 1% BME) and sealed with nail polish. Cells were imaged on a Zeiss Eclipse Ti microscope equipped with a 150 mW 561 nm laser and a 300 mW 647 laser. For imaging, lasers intensities were set to achieve 50 mW at the rear lens of the objective. Illumination was applied at a sub-TIRF angle through the objective to improve the signal to noise ratio. Videos of 24,000 frames (12,000 frames per channel) were acquired by groups of 6 consecutive frames using the NIS Elements software (Nikon). Images were aligned using 100 nm Tetraspeck beads (Thermo Fisher). This software was also used for peak detection and image reconstruction. The localization of the EEA1 termini could be distorted a maximum of approximately 20 nm owing to the size of the antibodies. The localization accuracy of the secondary antibody was ~25 nm. Measured distances were determined in Fiji and represent distances between respective centres-of-mass. Representative experiment is shown, $n = 3$.

Sample preparation for optical trap experiments. Bead-supported membranes were prepared as described. The concentrations used were as in the microscopy experiments: 1 nM Rabex-5 (131–394), 100 nM Rab5-6×His and EEA1 concentrations (between 30 and 400 nM). Most experiments were performed at 40 nM EEA1, with additional trials taking place at 4 and 400 nM. At lowest concentrations, single transient events became difficult to observe (<5% had interactions). At the highest concentrations, events were often non-transient or repeated.

Electron microscopy. Samples were rotary-shadowed essentially as described³⁹. Briefly, samples were diluted in a spraying buffer, consisting of 100 mM ammonium acetate and 30% glycerol. Diluted samples were sprayed via a capillary onto freshly cleaved mica chips. These mica chips were mounted in the high vacuum evaporator (MED 020, Baltec) and dried. Specimens were platinum coated (5–7.5 nm) and carbon was evaporated. Following deposition, the replica was floated off and examined at 71,000× magnification and imaged onto a CCD (Morgagni 268D, FEI; Morada G2, Olympus).

Analysis of electron microscopy. Images obtained were processed in ImageJ by skeletonizing the particles. Lengths were determined directly from these data and represent an overestimation due to the granularity of the platinum shadowing (5–7.5 nm granules). The bouquet plots were generated by aligning the initial five segments of the molecules and the entire population set was plotted.

To determine the curvature measure, we first took the skeletonized curves and smoothed them with a window of 8.2 nm. These curves were then segmented with 301 equally spaced points, and these smoothed curves were used for the curvature calculation. We first attempted to define curvature at one segment length (~0.75 nm) but this analysis was too noisy to obtain meaningful description of the curves. We therefore determined the curvature by taking the difference of the tangents and dividing it by the arc length at a distance of ~15 nm (20 points). The variance of this measure was determined, and bootstrapping with resampling was used to determine errors over the whole population and for 1,000 iterations.

Although proteins are not homogeneous polymers, the WLC model captures essential aspects of the physics underlying their shape fluctuations^{40,41}. Calculation of fits to all mean tangent-correlations and the equilibration analysis were performed using Easyworm source code in Matlab⁴². First, the original skeletonized curves were segmented with 301 equally spaced points. These data were then used to calculate the tangent-correlations and the kurtosis plots. We fitted the regime whereby the kurtosis measurement defined that the molecules were equilibrated^{18,43,44}. This distance therefore varied (see Extended Data Fig. 6, kurtosis plots), but the estimation of persistence length was only weakly dependent on this distance. The fitting routines were then implemented up to the thermal equilibration distance with bootstrapping with resampling, which was run for the whole population and 1,000 times to obtain errors. These are given as mean ± standard deviation. For values and fit statistics, please refer to Supplementary Data Table. We did not apply the WLC model to the swapped mutant (Extended Data Fig. 4h) because of the lack of significant structural changes upon Rab5 binding (Fig. 2f and Extended Data Fig. 4f).

The analytical fitting to the radial distribution functions was performed in Python¹⁸. The radial distribution function for a worm-like chain is the probability density for finding the end points of the polymer. The polymers are considered as embedded in a two-dimensional space in this scheme. This treatment adopts the continuum model of the polymer, thereby defining the statistical properties via free energy calculation. Fitting to analytical solution of the WLC yielded a

mean effective persistence length of 270 ± 14 nm for EEA1 alone (mean ± error of fit), and two populations of effective persistence lengths (26 ± 2 nm (67%) and 300 ± 14 nm (33%)) for EEA1 in the presence of Rab5-GTP-γS.

Optical tweezer experiments. A custom-built high-resolution dual-trap optical tweezer microscope was used^{45,46}. A single stable solid-state laser (Spectra-Physics, 5 W) was split by polarization into two traps that could be independently manoeuvred. Forces were measured independently in both traps by back-focal plane interferometry. Absolute distances between the two traps were determined by template-based video microscopy analysis (43 ± 2 nm per pixel) and offset-corrected for each microsphere pair by repeatedly contacting the microspheres after each experiment. The template detection algorithm had subpixel accuracy, at an estimated uncertainty in absolute distance measurements to be not more than ±20 nm. Bead displacement was calculated according to $\Delta F = -\kappa\Delta y$. Extended Data Fig. 7g demonstrates the sensitivity of the instrument via the Allan deviation⁴⁷ for averaging times greater than 100 ms.

All optical tweezer experiments were performed with 2 µm silica size-standard microspheres (Corpuscular), at a temperature of 26 ± 2 °C in a laminar flow chamber with buffers containing 35% glycerol to prevent sedimentation of the silica microspheres. Thermal calibration of the optical traps was performed with the power spectrum method using a dynamic viscosity of 3.1 mPas (ref. 48) (mean trap stiffness: trap 1, $\kappa_1 = 0.035 \pm 0.007$ pN/nm; trap 2, $\kappa_2 = 0.029 \pm 0.007$ pN/nm), leading to an overall trap stiffness of $\kappa_T = 0.0159$ pN/nm (yellow response curve in Extended Data Fig. 7h). Data were acquired at 1 kHz and further processed using custom-written software in R. Spurious electronic noise at 50 Hz was filtered using a fifth-order Butterworth notch filter from 49 to 51 Hz.

For probing the interactions of EEA1 with Rab5 without any assumptions on the shape of EEA1, a distance agnostic protocol with consecutive cycles of approaching, waiting (20 s) and retraction was used, approaching closer in each iteration (Fig. 3b). The stationary segments were then subjected to automatic change-point analysis to identify regions of the time series longer than 100 ms with significantly different mean and variance⁴⁹. Events thus identified were classified as transient if the mean and variance went back to base levels within the stationary segment (see examples in force traces in Fig. 3c and Extended Data Fig. 7). Mean times of interactions were 3.4 ± 0.6 s for GTP-γS and 0.9 ± 0.2 s for GTP. A fluctuation analysis of the differential distance signal during these events gave an estimated tether misalignment of less than 30° in all interactions⁵⁰. Only transient events were further processed. Silica beads alone as a negative control measured a mean contact distance of 22 nm (Fig. 3d, grey).

To calculate the persistence length for individual captured molecules we determined the equilibrium extension, z_{eq} , from the capture distance D (nm), the average measured force increase upon tethering ΔF (pN) and the known displacements from each trap $\Delta x_1 = \Delta F/\kappa_1$ and $\Delta x_2 = \Delta F/\kappa_2$ as $z_{eq} = D - \Delta x_1 - \Delta x_2$. With this distance, the persistence length was calculated according to⁵¹

$$\lambda(\Delta F, z_{eq}) = \frac{k_B T}{\Delta F} \left(\frac{z_{eq}}{L} - \frac{1}{4} + \frac{1}{4(1 - z_{eq}/L)^2} \right)$$

Similarly, to estimate the magnitude of the entropic collapse force, this formula was applied to the equilibrium extensions of EEA1, as estimated by the end-to-end distances of the molecules from electron microscopy. Values determined were (median and bounds at (2.5%, 97.5%)) EEA1, 23 (14, 33) nm; extended, 73 (60, 88) nm; swapped, 26 (21, 30) nm; 10×His, 78 (35, 140) nm. Values reported are medians and 95% confidence intervals determined from bootstrapping.

Generation of HeLa EEA1-KO cell line. HeLa EEA1-KO lines were generated using CRISPR-Cas9 technology⁵² on HeLa-Kyoto cell lines obtained from the BAC recombineering facility at the Max Planck Institute of Molecular Cell Biology and Genetics. Cell lines were tested for mycoplasma and authenticated (Multiplexion, Heidelberg). pSpCas9(BB-2A-GFP (PX458) and pSpCas9(BB)-2A-Puro (PX459) were a gift from F. Zhang (Addgene plasmid 48138, 48139). A PX458 plasmid encoding a GFP-labelled Cas9 nuclease and the sgRNA sequence (from GECKO⁵² library 17446, GTGGTTAAACCATGTTAAGG, targeting first exon) was transfected into standard HeLa Kyoto cells with Lipofectamine 2000 following the manufacturer's instructions. Cells were cultured in DMEM media supplemented with 10% FBS and 1% penicillin-streptomycin at 37 °C and 5% CO₂. After 3 days, the transfected cells were FACS sorted by their GFP fluorescence into 96-well plates to obtain single clones and visually inspected⁵³. These clones were then screened by western blotting and in-del formation confirmed sequencing of genomic DNA (primer forward, AGCGCCGTCGCCACCG; reverse, TAAGCGCCTGCCGGGCTG). Note the region is extremely GC-rich (75%, ±250 nt from targeted indel region). Additionally, a mixed-clonal line was obtained by transfection of HeLa Kyoto with PX459 with the above sgRNA sequence. After 72 h from transfection, cells were exchanged into media supplemented with 0.5 µg/ml puromycin (concentration determined in separated

experiment) and selected for 3 days. All imaging experiments were confirmed on this secondary line.

Endocytosis rescue assays. Wild-type EEA1 and the extended and swapped variants (Extended Data Fig. 3) were cloned into customized mammalian expression plasmids under the CMV promoter resulting in untagged proteins. HeLa or HeLa EEA1-KO cells were seeded into 96-well plates and transfected (or mock transfected) after 48 h. Following 48 h after transfection, cells were exchanged into serum-free media containing 8.2 µg/ml LDL-Alexa 488 (prepared as previously described¹⁶) or 100 ng/ml EGF-Alexa 488 (E13345, Thermo Fisher) for 10 min at 37 °C, and washed in PBS then fixed in 4% paraformaldehyde.

Automated confocal immunofluorescence microscopy and analysis. Fixed cells were stained with antibodies against EEA1 (laboratory-made rabbit) and Rab5 (610724, prepared in mouse, BD Biosciences) as previously described²⁴. DAPI was used to stain the nuclei. Not all early endosomes harbour EEA1 (ref. 54) and other tethering factors could compensate for EEA1 (refs 24, 55). All imaging was performed on a Yokogawa CV7000 s automated spinning disc confocal using a 60 × 1.2 numerical aperture objective. Fifteen images were acquired per well and each condition was duplicated at least twice per plate, resulting in 30 or more images per condition.

Image analysis used home-made software, MotionTracking, as previously described^{56,57}. Images were first corrected for illumination, chromatic aberration and physical shift using multicolour beads. All cells, nuclei and cell objects in corrected images were then segmented and their size, content and complexity calculated. The intensity of EEA1 in wild-type HeLa cells was measured to determine a wild-type intensity distribution. In the rescue experiments, an intensity threshold for the transfections was set at about two times the mean of wild-type cells (Extended Data Fig. 8i). Experiments were repeated at different seeding densities with similar results. Given a cell density threshold between 10 and 100 per image, we obtained an average of more than 300 cells per condition after filtering for the transfection level of EEA1, and more than 15,000 endosomes per experiment. A two-tailed *t*-test was used for significance calculations.

Cell electron microscopy. Cells in 3 cm diameter plastic dishes were processed for electron microscopy using a method⁵⁸ to provide particularly heavy staining of cellular components. Briefly, cells were fixed by addition of 2.5% glutaraldehyde in PBS for 1 h at room temperature and then washed with PBS. The cells were then processed as described⁵⁸ with sequential incubations in solutions containing potassium ferricyanide/osmium tetroxide, thiocarbonylhydrazide, osmium tetroxide, uranyl acetate and lead nitrate in aspartic acid before dehydration and flat embedding in resin. Sections were cut parallel to the substratum and analysed unstained in a JEOL 1011 transmission electron microscope (Tokyo, Japan). Images for quantitation were collected from coded samples (double blind) to avoid bias.

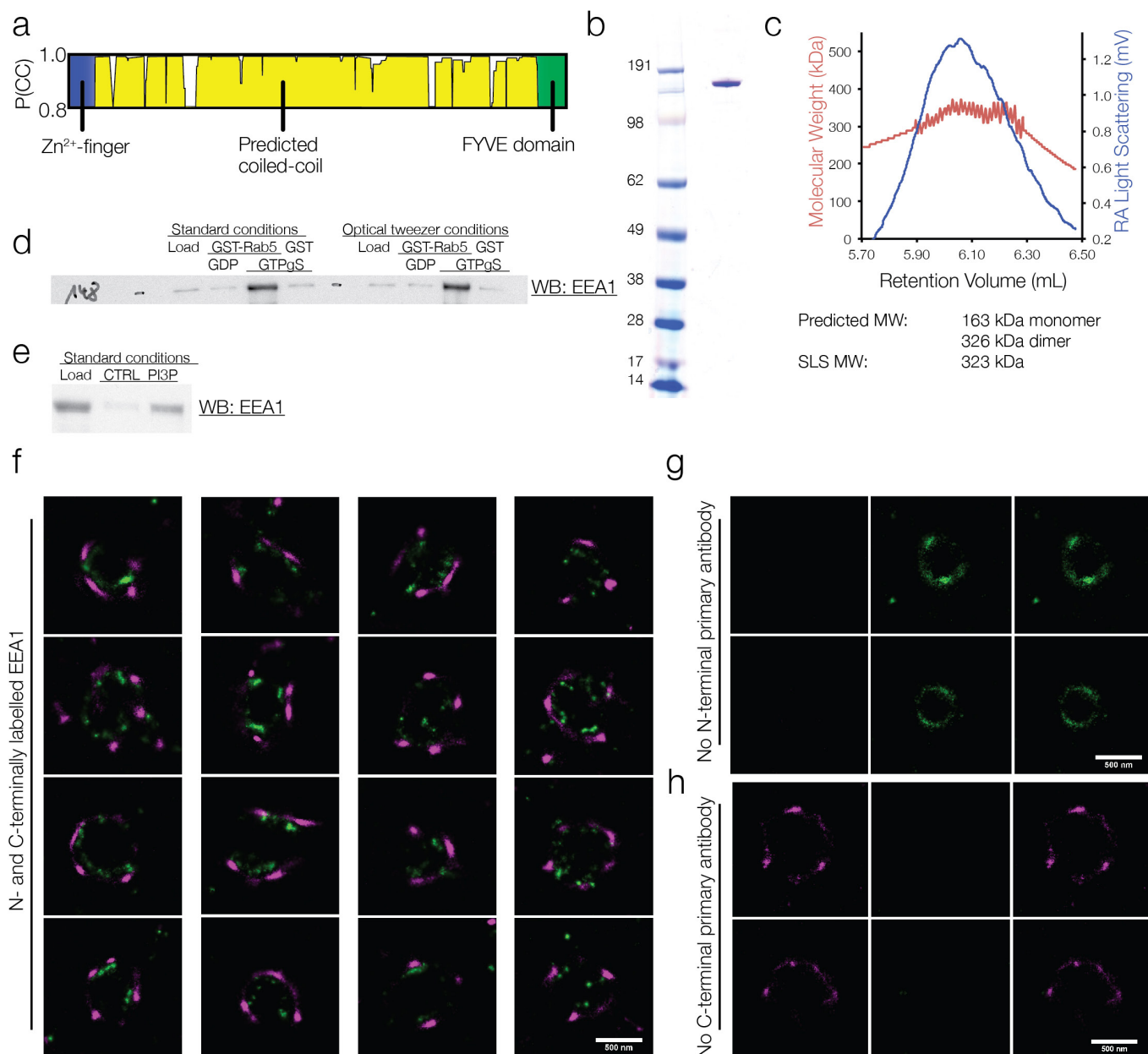
Distance analysis used ImageJ. To correct for thickness of slices (60 nm), the following equation was used:

$$P(R) = \frac{1}{Z} \int_0^H P_0(\sqrt{R^2 - h^2}) \frac{R}{\sqrt{R^2 - h^2}} dh,$$

where $P_0(r)$ is the apparent 2D distance distribution, R is the 3D distance, H is the thickness of the slice and Z is the normalization constant. Uncorrected distance was measured at 119.8 ± 78.2 nm (mean \pm s.d.), which resulted in 130.0 ± 76.8 nm corrected.

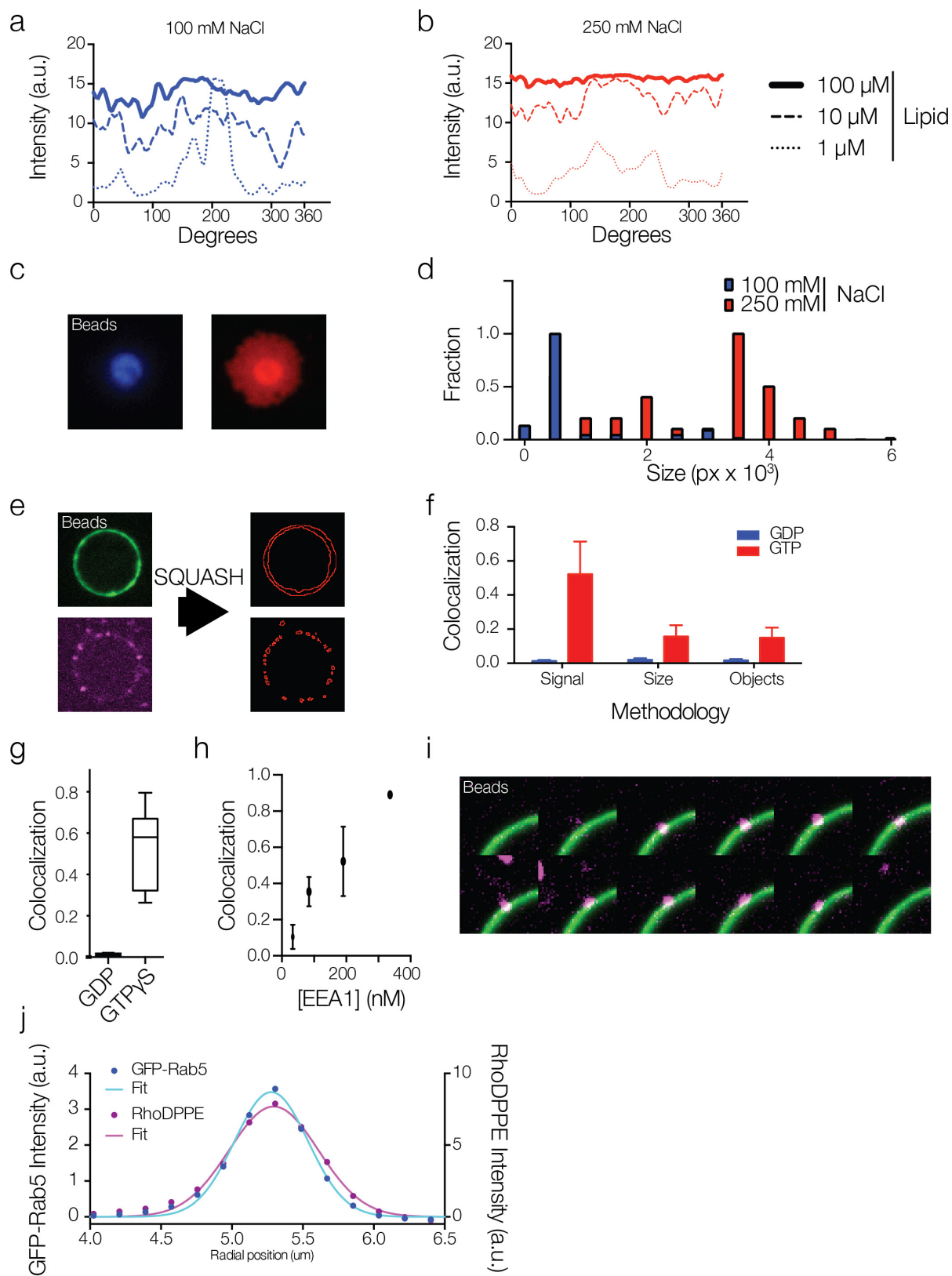
30. Sheffield, P., Garrard, S. & Derewenda, Z. Overcoming expression and purification problems of RhoGDI using a family of “parallel” expression vectors. *Protein Expr. Purif.* **15**, 34–39 (1999).
31. Delprato, A., Merithew, E. & Lambright, D. G. Structure, exchange determinants, and family-wide rab specificity of the tandem helical bundle and Vps9 domains of Rabex-5. *Cell* **118**, 607–617 (2004).

32. Horiuchi, H. *et al.* A novel Rab5 GDP/GTP exchange factor complexed to Rabaptin-5 links nucleotide exchange to effector recruitment and function. *Cell* **90**, 1149–1159 (1997).
33. Boura, E. & Hurley, J. H. Structural basis for membrane targeting by the MVB12-associated β -prism domain of the human ESCRT-I MVB12 subunit. *Proc. Natl Acad. Sci. USA* **109**, 1901–1906 (2012).
34. Murray, D. H., Tamm, L. K. & Kiessling, V. Supported double membranes. *J. Struct. Biol.* **168**, 183–189 (2009).
35. Neumann, S., Pucadyil, T. J. & Schmid, S. L. Analyzing membrane remodeling and fission using supported bilayers with excess membrane reservoir. *Nature Protocols* **8**, 213–222 (2013).
36. Pucadyil, T. J. & Schmid, S. L. Real-time visualization of dynamin-catalyzed membrane fission and vesicle release. *Cell* **135**, 1263–1275 (2008).
37. Rizk, A. *et al.* Segmentation and quantification of subcellular structures in fluorescence microscopy images using Squassh. *Nature Protocols* **9**, 586–596 (2014).
38. Lo, S. Y. *et al.* Intrinsic tethering activity of endosomal Rab proteins. *Nature Struct. Mol. Biol.* **19**, 40–47 (2011).
39. Tyler, J. M. & Branton, D. Rotary shadowing of extended molecules dried from glycerol. *J. Ultrastruct. Res.* **71**, 95–102 (1980).
40. Gittes, F., Mickey, B., Nettleton, J. & Howard, J. Flexural rigidity of microtubules and actin filaments measured from thermal fluctuations in shape. *J. Cell Biol.* **120**, 923–934 (1993).
41. Eeftens, J. M. *et al.* Condensin Smc2-Smc4 dimers are flexible and dynamic. *Cell Reports* **14**, 1813–1818 (2016).
42. Lamour, G., Kirkegaard, J. B., Li, H., Knowles, T. P. & Gsponer, J. Easyworm: an open-source software tool to determine the mechanical properties of worm-like chains. *Source Code Biol. Med.* **9**, 16 (2014).
43. Rivetti, C., Guthold, M. & Bustamante, C. Scanning force microscopy of DNA deposited onto mica: equilibration versus kinetic trapping studied by statistical polymer chain analysis. *J. Mol. Biol.* **264**, 919–932 (1996).
44. Valle, F., Favre, M., De Los Rios, P., Rosa, A. & Dietler, G. Scaling exponents and probability distributions of DNA end-to-end distance. *Phys. Rev. Lett.* **95**, 158105 (2005).
45. Lisica, A. *et al.* Mechanisms of backtrack recovery by RNA polymerases I and II. *Proc. Natl Acad. Sci. USA* **113**, 2946–2951 (2016).
46. Jahnel, M., Behrndt, M., Jannasch, A., Schäffer, E. & Grill, S. W. Measuring the complete force field of an optical trap. *Opt. Lett.* **36**, 1260–1262 (2011).
47. Czerwinski, F., Richardson, A. C. & Oddershede, L. B. Quantifying noise in optical tweezers by allan variance. *Opt. Express* **17**, 13255–13269 (2009).
48. Nørrelykke, S. F. & Flyvbjerg, H. Power spectrum analysis with least-squares fitting: amplitude bias and its elimination, with application to optical tweezers and atomic force microscope cantilevers. *Rev. Sci. Instrum.* **81**, 075103 (2010).
49. Killick, R., Fearnhead, P. & Eckley, I. A. Optimal detection of changepoints with a linear computational cost. *J. Am. Stat. Assoc.* **107**, 1590–1598 (2012).
50. Ribezzi-Crivellari, M. & Ritort, F. Force spectroscopy with dual-trap optical tweezers: molecular stiffness measurements and coupled fluctuations analysis. *Biophys. J.* **103**, 1919–1928 (2012).
51. Marko, J. F. & Siggia, E. D. Statistical mechanics of supercoiled DNA. *Phys. Rev. E* **52**, 2912–2938 (1995).
52. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8**, 2281–2308 (2013).
53. Poser, I. *et al.* BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nature Methods* **5**, 409–415 (2008).
54. Kalaidzidis, I. *et al.* APPL endosomes are not obligatory endocytic intermediates but act as stable cargo-sorting compartments. *J. Cell Biol.* **211**, 123–144 (2015).
55. Peplowska, K., Markgraf, D. F., Ostrowicz, C. W., Bange, G. & Ungermann, C. The CORVET tethering complex interacts with the yeast Rab5 homolog Vps21 and is involved in endo-lysosomal biogenesis. *Dev. Cell* **12**, 739–750 (2007).
56. Collinet, C. *et al.* Systems survey of endocytosis by multiparametric image analysis. *Nature* **464**, 243–249 (2010).
57. Gilleron, J. *et al.* Image-based analysis of lipid nanoparticle-mediated siRNA delivery, intracellular trafficking and endosomal escape. *Nature Biotechnol.* **31**, 638–646 (2013).
58. Takasato, M. *et al.* Kidney organoids from human iPS cells contain multiple lineages and model human nephrogenesis. *Nature* **526**, 564–568 (2015).



Extended Data Figure 1 | EEA1 is a predicted extended coiled-coil dimer that binds Rab5 in a GTP-dependent manner and extends outwards from endosomes **a**, Human EEA1 in COILS prediction reveals a clear coiled-structure flanked by the Rab5-binding Zn²⁺-finger on the N terminus and PI(3)P binding FYVE domain on the C terminus. **b**, Coomassie-stained gel of human EEA1 expressed as a GST fusion in SF+ insect cells and purified by GS affinity, cleaved on resin, and subsequently concentrated and separated from smaller contaminants by size-exclusion chromatography on a Superose 6 column. **c**, Static light scattering in line with size-exclusion chromatography reveals a molecular mass of 323 kDa, compared with a theoretical molecular mass of 326 kDa for a dimeric protein. **d**, Purified protein binds Rab5 in both standard and optical tweezer conditions (35% glycerol) in a GTP-dependent manner. GST or GST-Rab5 was purified and conjugated to GS resin, and subsequently nucleotide was exchanged to either GTP- γ S or GDP using

EDTA-Mg²⁺-mediated exchange and subsequent wash. The GST resin was then incubated with EEA1 in either the standard or optical tweezers buffer, washed three times, and beads were then blotted for EEA1. **e**, Recombinant EEA1 binds specifically to PI(3)P liposomes. When mixed with POPC:POPS 85:15 liposomes, no EEA1 is observed in the liposome pellet (CTRL). In contrast, EEA1 is pelleted with control POPC:POPS:PI(3)P 80:15:5 liposomes (PI3P). **f**, The N-terminal Zn²⁺-finger and C-terminal FYVE domain of EEA1 were differentially labelled with specific antibodies and STORM microscopy performed to define their localization in HeLa cells. Representative STORM images of EEA1 radial extension from endosome of *n* = 22. Scale bar, 500 nm. **g, h**, Primary antibody binding controls for N and C termini. Primary antibodies for the N (**g**) and C (**h**) termini were left out of the staining, resulting in no unspecific secondary staining for each. Representative of *n* = 5. Scale bar, 500 nm.



Extended Data Figure 2 | See next page for caption.

Extended Data Figure 2 | Validation of bead-supported lipid bilayers for optical tweezers, and bead tethering experiment controls and methods.

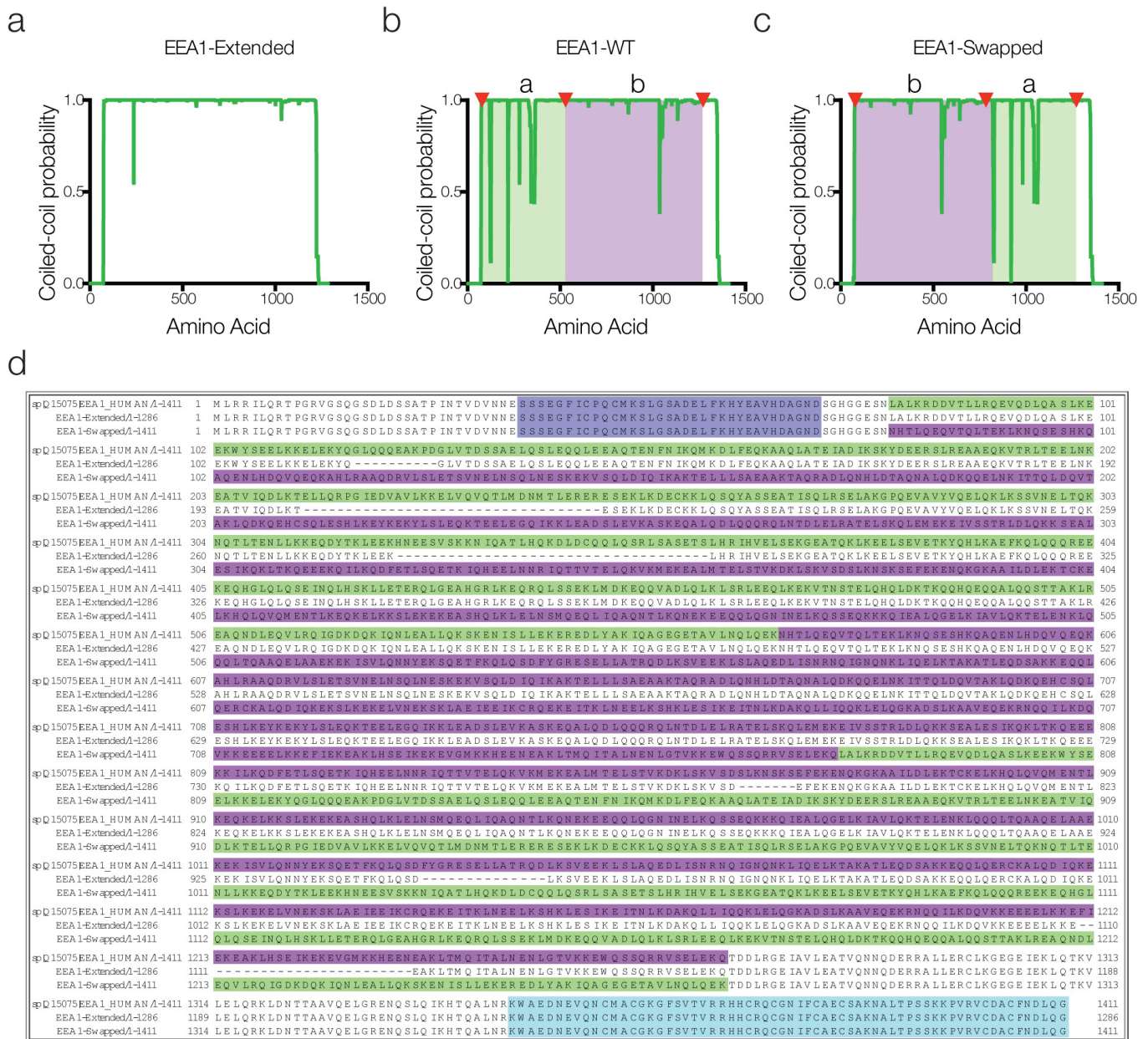
To optimize the conditions for forming supported lipid bilayers on the 2–10 μm beads, we systematically investigated the dependence of membrane formation on salt and liposome concentration. **a**, Fluorescent profiles of supported lipid bilayer bead cross sections. At high liposome concentration (100 μM , solid line) during formation of the bilayer on the silica bead, the bead-supported membrane fluorescence intensity is circumferentially homogenous. At lower lipid concentrations (10 and 1 μM , dashed and dotted lines), less than full coverage is achieved and the supported bilayer is inhomogeneous. **b**, Consistent with previous reports, increasing salt concentrations result in more homogenous membrane coverage. **c**, Representative examples of the ‘spilled-out’ membrane of beads prepared at 100 mM (top, blue) and 250 mM (bottom, red) NaCl salt and 100 μM liposomes, of $n = 5$. **d**, Histogram of the size of membrane spilled from the beads onto the substrate when prepared at 100 and 250 mM NaCl (blue and red, respectively). This indicated that the lower salt samples (blue) were homogeneously covered with membrane and that they had little excess present, and therefore the optimal conditions for formation of membrane on the silica beads used in tethering and in optical tweezer experiments. **e**, Segmentation of beads and vesicles by

the SQUASH method. Bead-supported bilayers and vesicles (green and magenta, respectively) were segmented as illustrated by red outlines to determine their co-localization. Representative of $n = 1$ generated for schematic. **f**, Methodology comparison for co-localization in GDP and GTP- γS conditions. All methods give $P < 0.01$ in a two-tailed Student's t -test. Co-localization by signal is better than by size or object, as vesicles become undercounted at high concentrations. Mean \pm s.d., $n = 5$.

g, Co-localization of liposomes (PI(3)P, magenta) to the bead-supported membrane (GFP-Rab5, green) was strictly dependent on GTP- γS .

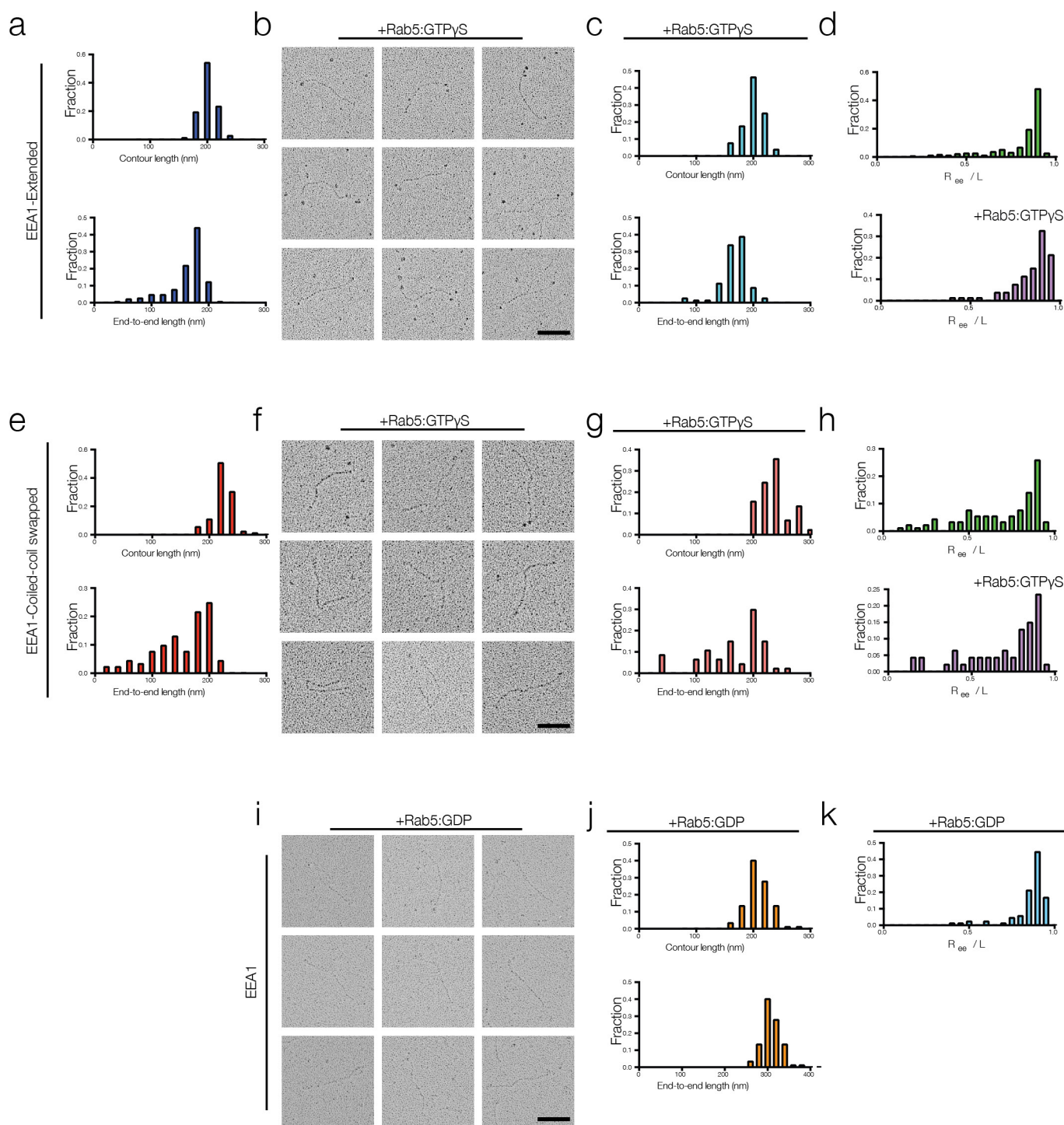
Box-whisker plot with minimum/maximum error, $n = 5$. **h**, The co-localization of liposomes to the supported membrane was dependent on EEA1 concentration. At higher concentrations of EEA1, co-localization approached 100%. These concentrations are within the range of the concentration of endogenous protein²³. Mean \pm s.d., $n = 5$. **i**, Time-lapse micrographs of the bead-supported bilayer labelled with GFP-Rab5 (green), and a dynamically tethered vesicle (magenta). Vesicles were observed to tether and reversibly leave the membrane, as well as diffuse about its surface. Images displayed were acquired at 350 ms intervals as z -stacks. Representative of $n = 1$ to acquire video. Scale bar, 2 μm .

j, Example fits for radial line-profile data.



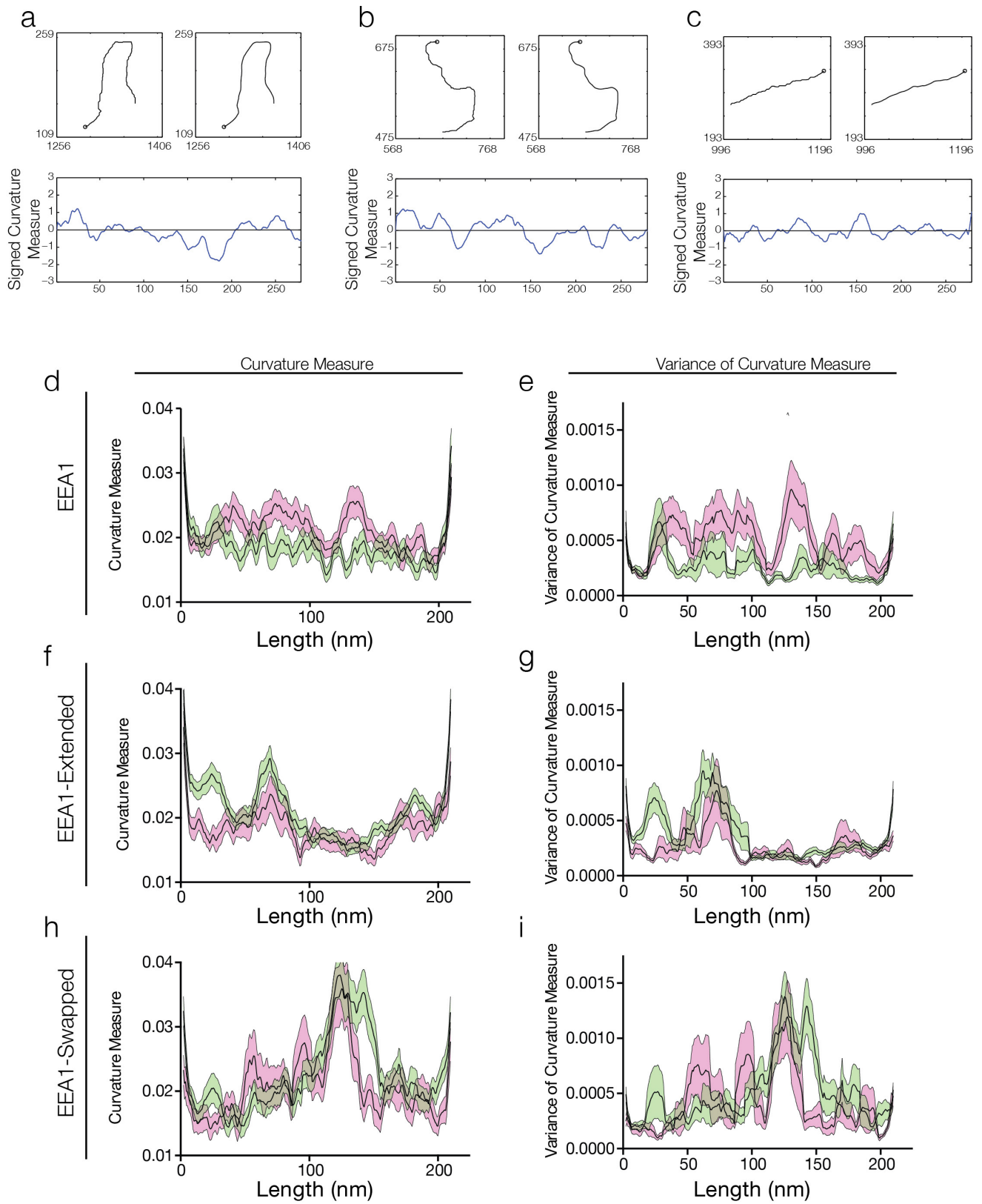
Extended Data Figure 3 | Structure prediction and sequence description of EEA1 mutants. **a**, COILS prediction for extended EEA1 mutant, revealing removal of most of the discontinuities in the coiled-coil. **b**, **c**, The swapped EEA1 mutant has a rearranged coiled-coil. The coiled-coil was split as indicated by red triangles in the original EEA1-WT (**b**), and the two regions *a* (shaded green) and *b* (shaded magenta) were rearranged in a synthetic gene, producing the swapped EEA1 variant maintaining the features and sequence of the original coiled-coil, but in an alternative

location (**c**). **d**, Full sequence alignment for human EEA1 and the extended and swapped mutants used in the study. The crystal structure (Protein Data Bank accession number 3MJH) for the Zn²⁺-finger domain is marked in dark blue close to the N terminus. Segment *a* of the coiled-coil region is marked in green, and segment *b* in magenta. The crystal structure (Protein Data Bank accession number 1JOC) of the C-terminal FYVE domain and portion of the coiled-coil is marked in cyan. Details of the mutant constructs are found in the Methods.



Extended Data Figure 4 | Extended and swapped EEA1 mutants exhibit limited changes in the presence of Rab5:GTP- γ S. **a, e,** Rotary-shadowed EEA1-extended particles and EEA1-swapped mutants were skeletonized and analysed in ImageJ for contour length (top), resulting in normally distributed contour length histograms. The end-to-end length histograms (bottom) are similarly distributed. These data were collected on N-terminally MBP-tagged samples. Compare with wild-type in Fig. 2b, d; $n = 212$ for the extended and $n = 93$ for the swapped variants. **b–d, f, g,** The EEA1 mutants revealed limited changes to their curvature

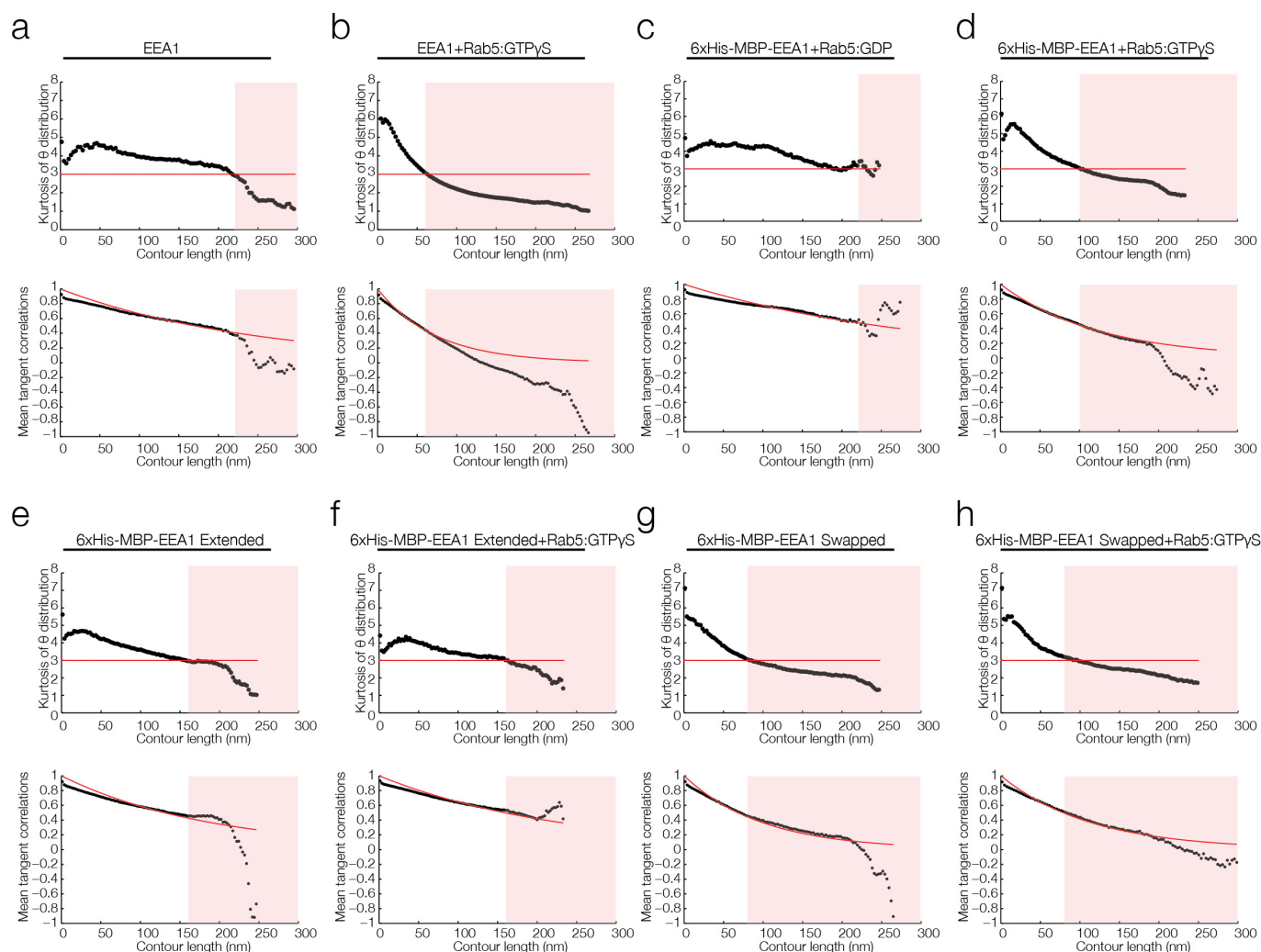
in the presence of Rab5:GTP- γ S (**b, f**; compare Fig. 2i, j), and therefore minor changes to their contour and end-to-end length histograms (**c, g**) and radial distribution plots (**d, h**); $n = 80$ for the extended and $n = 47$ for the swapped variants. **i, j,** Rotary-shadowing electron microscopy of EEA1 in the presence of Rab5:GDP ($n = 90$), N-terminally MBP-tagged, revealed no change in appearance compared with the absence of Rab5 entirely (Fig. 2a), and no effect of N-terminal tagging relative to wild-type EEA1. **k,** Radial distribution function of EEA1 in the presence of Rab5:GDP (compare **d, h**; Fig. 2g); $n = 90$.



Extended Data Figure 5 | See next page for caption.

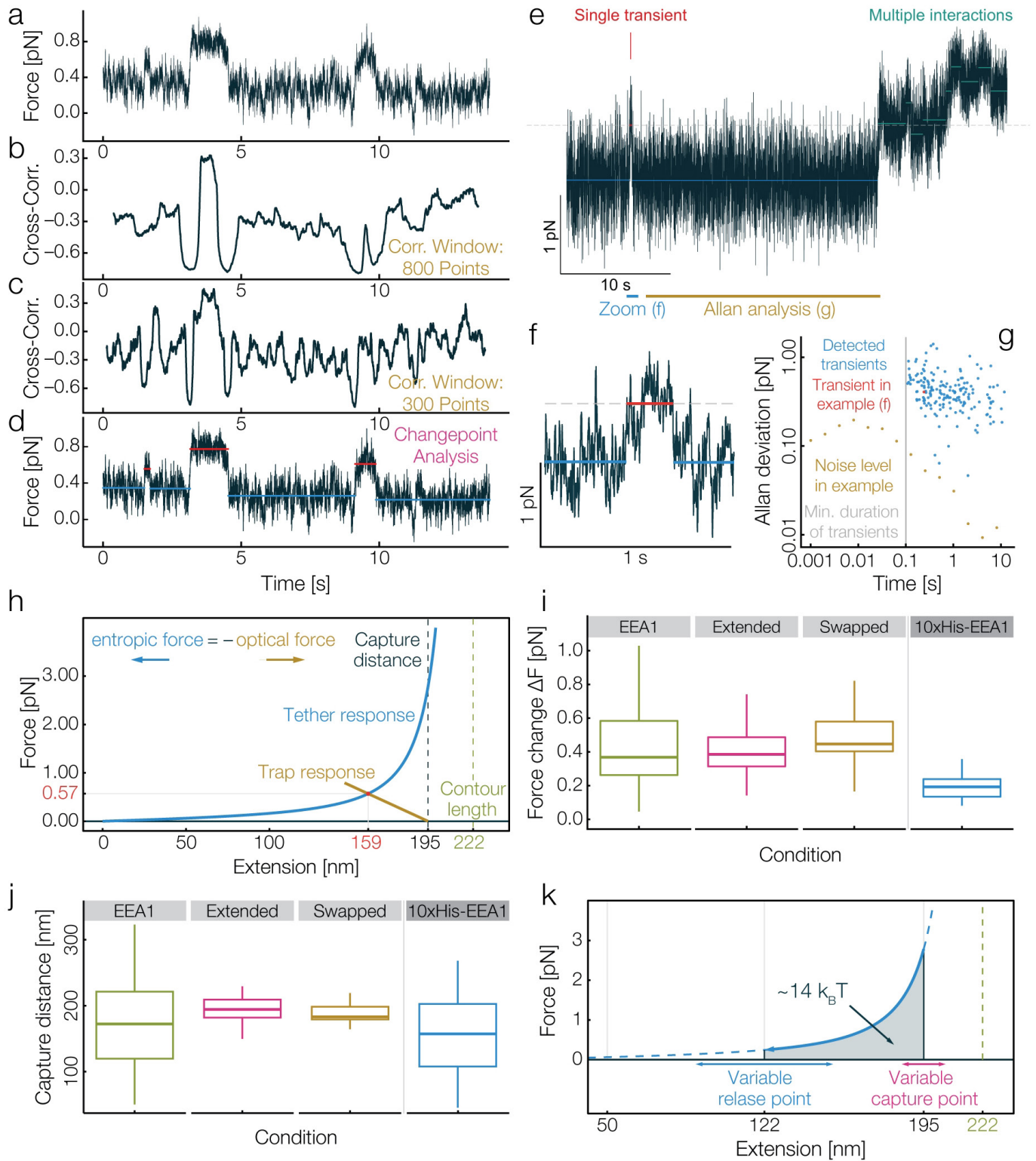
Extended Data Figure 5 | Representative segmentation, smoothing and signed curvature measures for EEA1, and averages for EEA1 and mutants. EEA1 and EEA1 mutants were skeletonized and smoothed using a moving average filter with a window of 8.2 nm, segmented to 300 equally spaced segments and aligned N terminus to C terminus by recognition of an N-terminal MBP-tag. Their curvature was calculated at 15 nm distances along the length of the proteins and plotted. **a–c**, Representative examples of rotary shadowing derived EEA1 curves. The original data appear in the first panel, with the second panel revealing the data after smoothing for comparison (Methods). The curvature measure, determined by how the tangents to the contour change at a distance of 15 nm along the contour is plotted below. Note that the choice of sign for the curvature measure is arbitrary for each molecule. **d, e**, Curvature measure and

variance of this measure for EEA1 in the presence of Rab5:GDP (green) and EEA1 in the presence of Rab5:GTP- γ S (magenta); $n = 90$, $n = 145$, respectively. Alignment of EEA1 curvature from the electron microscopy data reveals an increase in curvature over the length of the molecule upon Rab5 binding, whereas the extended and swapped EEA1 variants show no change. All curvature values were taken to be positive given that the N-terminal MBP could be recognized but the handedness of the molecule adsorbed to the grid could not be inferred. Bootstrapping with resampling at full population size was performed for 1,000 iterations to determine errors. **f, g**, Extended EEA1 variant in the absence (green) and in the presence of Rab5:GTP- γ S (magenta); $n = 212$, $n = 80$, respectively. **h, i**, Swapped EEA1 variant in the absence (green) and in the presence of Rab5:GTP- γ S (magenta); $n = 93$, $n = 47$, respectively.



Extended Data Figure 6 | Detailed persistence length and equilibration analysis for EEA1 and variants. To validate the methodology used for analysis of the persistence lengths, and to assure internal consistency in analysis methods, we systematically applied the analysis to EEA1 (and mutants, see Supplementary Data Table). The skeletonized curves were segmented to 300 equally spaced segments, where θ describes the angle between segments. The tangent–tangent correlations were then determined for the entire ensembles. **a–h**, To determine the molecular equilibration of EEA1 and variants from 3D to 2D, the kurtosis of the theta distribution (top) was calculated. Full equilibration to 2D gives a

value of 3.0, and for 3D the expected value is 1.8 as the angle distributions become Gaussian. As expected, the measured kurtosis is approximately 3.0 until lengths above the persistence length of the molecule, where the equilibration begins to fail. The value at which the kurtosis began to diverge from 2D was taken as the limit for subsequent measurements, as beyond this limit (red shaded region) 3D fluctuations are not retained and as such the consequences of surface adsorption are uncertain. Next, the tangent–tangent correlation was calculated across the ensemble and fitted up to the divergence of the kurtosis (red shaded region).

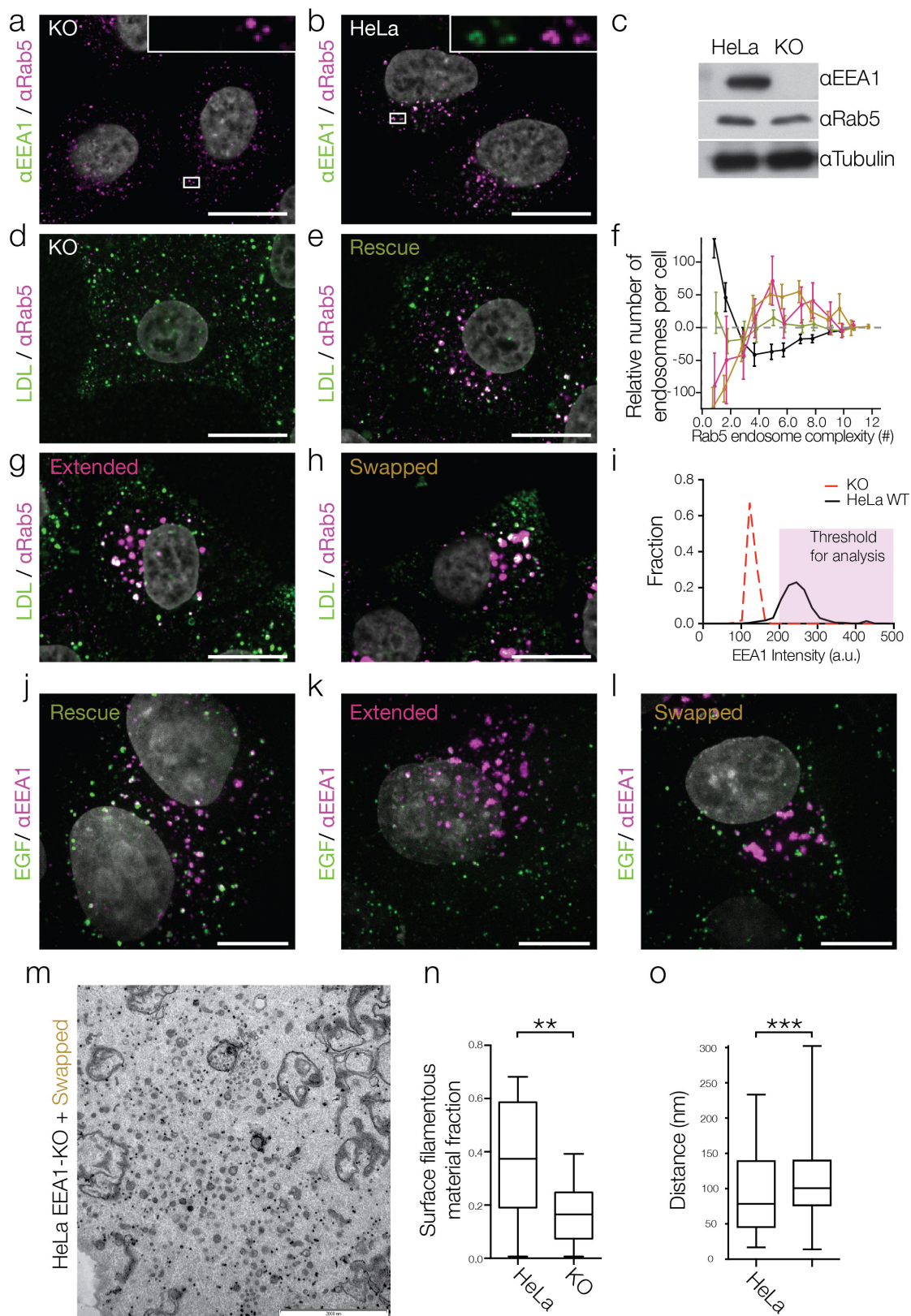


Extended Data Figure 7 | See next page for caption.

Extended Data Figure 7 | Supplementary data related to optical

tweezer experiments. a, Change-point analysis was used to identify changes in the mean and variance of the combined force signal. An example plot of averaged force (linear combination of signals from both traps) with respect to time. Data have been collected at 1 kHz. Two long transient interactions can be clearly identified. **b, c,** Cross-correlation of the force signals from each trap are not sufficient to reveal stepwise interactions as they are time-averaged. By applying cross-correlation over a correlation window of 0.8 s (**b**) or 0.3 s (**c**), long transient interactions (that is, at ~ 4 s) could be identified. However, an unbiased identification of short transients (that is, at ~ 9 s) by this method was not possible. All identified long transient interactions showed characteristic changes in the cross-correlation: anti-correlation as beads are pulled together, and correlation after tethering was established. **d,** Change-point analysis was used to detect both changes in mean and variance of the combined force signal, and thereby identify transient interactions (red line). This procedure has the additional advantage of defining clear boundaries to stepwise processes. **e,** The possibility of multiple tethers taking part in the reaction was observed. Averaged force trace for wild-type EEA1 occasionally showed signals consistent with multiple interactions (cyan), in addition to single transient interactions (red). **f,** Zoom into time series around the transient interaction identified in the previous panel. To a first approximation, the dynamic interactions were fitted as piecewise constant steps (red). Note also two very short (<10 ms) spikes of similar magnitude (to the left and right of identified interaction) occurred but are not used in further analysis. Only transients with a duration longer than 100 ms were analysed. **g,** To illustrate the sensitivity of the optical tweezer experiments, a noise analysis was performed on the segment outlined in the top panel (yellow, labelled Allan analysis). The Allan deviation (square root of Allan variance, in piconewtons) gives a threshold for detecting a signal change over different averaging windows. All detected transients (blue) are at minimum an order of magnitude above this threshold. To provide perspective, the transient in the above example is indicated as a red dot. **h,** The entropic collapse force is balanced in the tweezer experiments below its peak value. The balance between the average restoring force in the optical traps (brown) and the entropic collapse force of EEA1 (blue) in the bound state gives the measured equilibrium force

and extension (red dot). The schematic assumes the measured capture distance of 195 nm, a persistence length in the Rab5:GTP-bound state of $\lambda_b = 26$ nm, and a contour length of 222 nm. The overall trap response of the dual-trap system is treated as two springs in series with the mean trap stiffness in trap 1 ($\kappa_1 = 0.035 \pm 0.007$ pN/nm) and the mean trap stiffness in trap 2 ($\kappa_2 = 0.029 \pm 0.007$ pN/nm), leading to an overall trap stiffness of $\kappa_T = 0.0159$ pN/nm (brown line). Given these parameters, the predicted equilibrium force in the optical trap for Rab5-bound EEA1 is ~ 0.6 pN and the predicted equilibrium extension ~ 160 nm. **i,** Force changes upon capture for Rab5:GTP-bound EEA1 and the extended and swapped variants. Force was measured from change-point analysis for transient interactions between EEA1 beads and Rab5:GTP beads. To test binding per se, the force change for $10\times$ His-EEA1 beads tethered to Ni-NTA beads was similarly determined from established connections. For $10\times$ His-EEA1, no transient interactions could be observed. Median change in force and 95% confidence interval from bootstrapping with resampling (lower and upper bounds at (2.5%, 97.5%)) were determined. EEA1, 0.37 (0.31, 0.46) pN; extended, 0.39 (0.35, 0.42) pN; swapped, 0.45 (0.41, 0.56) pN; $10\times$ His, 0.19 (0.14, 0.22) pN. **j,** Capture distances defined at the proximal distance upon which transient interactions were observed for Rab5-bound EEA1 and the extended and swapped variants. Median capture distance and 95% confidence interval from bootstrapping with resampling (lower and upper bounds at (2.5%, 97.5%)) were determined. EEA1, 168 (141, 182) nm; extended, 195 (189, 199) nm; swapped, 183 (179, 189) nm; $10\times$ His, 157 (120, 196) nm; $n = 60, 93, 27, 24$ per condition respectively. **k,** Mechanical work is performed as the tether collapses. The mechanical work performed during the relaxation to the new equilibrium extension is the integral under the force–extension curve. The exact value of the extracted work depends both on the capture distance (the extension at the moment of persistence length change) and on the release distance (the extension at the moment when Rab5 unbinds). The uncertainties in these extensions are different for the two positions, reflecting the different longitudinal fluctuations of the rigid or the flexible tether ($\lambda_{\text{flexible}} = 26$ nm (blue arrows), $\lambda_{\text{rigid}} = 300$ nm (magenta arrows)). For example, for a relaxation between the capture distance, $d_{\text{capture}} \approx 195$ nm and the release extension, $d_{\text{release}} \approx 122$ nm, the extracted mechanical work is $W \approx 14 k_B T$.

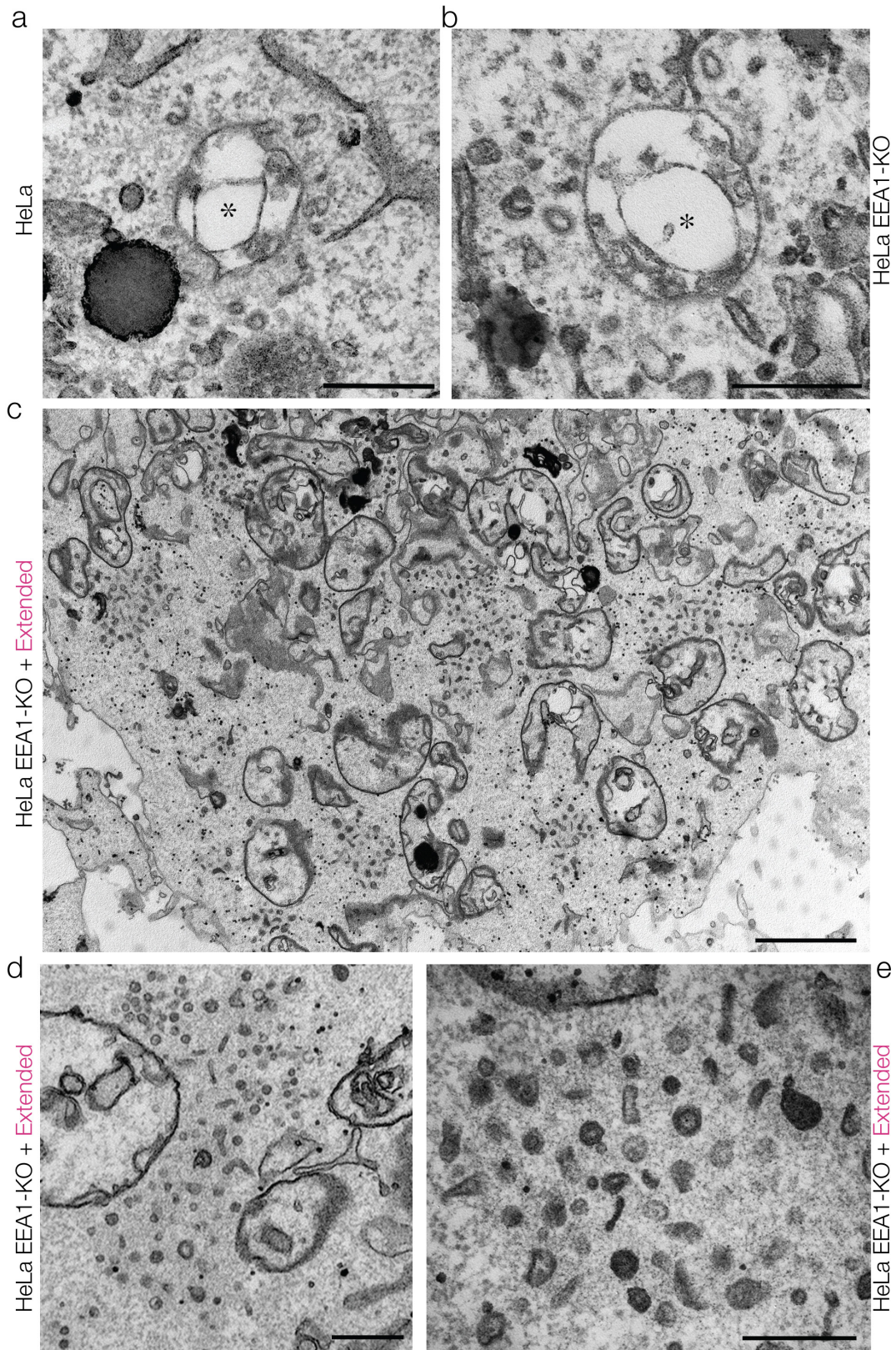


Extended Data Figure 8 | See next page for caption.

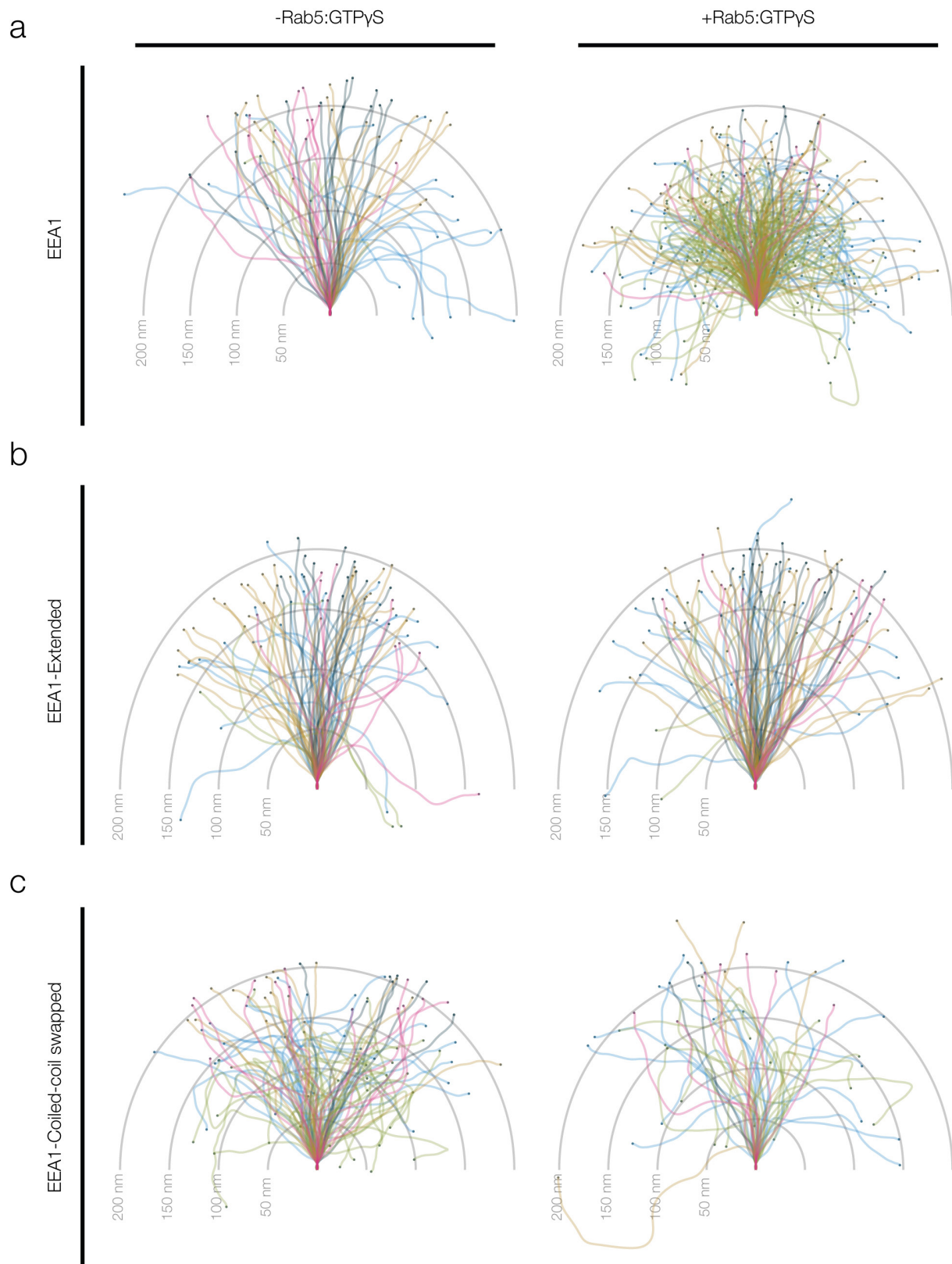
Extended Data Figure 8 | EEA1 mutants incapable of undergoing entropic collapse result in defects in endosomal trafficking.

a, b, Automated confocal immunofluorescence images ($n = 30$ each) of HeLa EEA1-KO and standard HeLa cells. EEA1 (green) and Rab5 (magenta). Scale bar, $10\ \mu\text{m}$. **c**, Western blot of HeLa and HeLa EEA1-KO clonal cell line for EEA1 and Rab5. **d, e, g, h**, Automated confocal images ($n = 30$ each) of HeLa EEA1-KO cells expressing no EEA1 (KO, **d**), rescued with wild-type EEA1 (rescue, **e**) or extended and swapped mutants (**g, h**). Cells were pulsed with fluorescently labelled cargo (LDL) (green) for 10 min, fixed and immunostained for Rab5 (magenta) and EEA1 (for EEA1, see Fig. 4). Magnified insets of endosomes are depicted at arrows. Scale bar, $10\ \mu\text{m}$. **f**, Relative complexity of Rab5 endosomes per cell. Each Rab5 endosome is segmented, and the segmented object requires a defined number of 2D Gaussian functions, hereby referred to as complexity. Relative to wild type, HeLa EEA1-KOs (black line) had a significantly reduced number of endosomes of high complexity (>3.0), but more endosomes defined simply by one or two Gaussian functions. Rescue experiments (red) revealed no significant difference in complexity. In contrast, both extended and swapped mutants (blue and green respectively) had significantly fewer simple endosomes of low complexity, and significantly more of higher complexity. Mean \pm s.d., $n = 30$.

i, Histogram of fluorescence intensity of EEA1 per cell. KO cell lines had a sharp peak of intensity at background levels, whereas wild-type HeLa cells had a normal distribution. Grey box represents threshold levels of EEA1 intensity per cell taken for analysis. **j–l**, EGF uptake experiments. Confocal images of HeLa EEA1-KOs expressing wild-type EEA1 (rescue, **j**) or extended and swapped mutants (**g, h**). Cells were pulsed with fluorescently labelled EGF (green) for 10 min, fixed and immunostained for EEA1 (magenta). Images shown are maximum intensity projections. Scale bar, $5\ \mu\text{m}$. **m**, HeLa EEA1-KO cells in which the swapped EEA1 mutant was reintroduced showed clusters of vesicles and more rarely the classical endosomal morphology. The clusters were clearly delineated by a zone of cytoplasm with a distinct density. Representative of $n = 19$. Scale bars, $2\ \mu\text{m}$. **n**, Further quantifications, and the swapped mutant ultrastructural phenotype. Fraction of endosomal surface containing filamentous material for HeLa and HeLa EEA1-KOs. Box-whisker plot with minimum/maximum values, $n = 22, 24$ endosomes. $**P < 0.01$, two-tailed Student's *t*-test. **o**, Distance measured between endosome and tethered vesicles (HeLa) or between vesicles within large clusters (extended) (surface-to-surface, $n = 158$ and 623 for HeLa and extended respectively; $***P < 10^{-4}$, two-tailed Student's *t*-test).



Extended Data Figure 9 | Unlabelled version of Fig. 5.



Extended Data Figure 10 | Bouquet plots of EEA1 and variants. EEA1 in the absence of Rab5 is predominantly extended. The initial five segments of the curves from rotary shadowing electron microscopy were aligned and the curves plotted with the end position highlighted (dots). Grey concentric hemispheres demarcate 50, 100, 150 and 200 nm extensions

from the origin. The end positions therefore resulted in a cloud of empirical positions for the EEA1 N terminus of EEA1 (left), and reveal the overall change in conformational space that can be occupied by EEA1 when bound to Rab5:GTP- γ S (right). **b**, Bouquet plots for the extended EEA1 variant. **c**, Bouquet plots for the swapped EEA1 variant.

Small molecule stabilization of the KSR inactive state antagonizes oncogenic Ras signalling

Neil S. Dhawan^{1,2*}, Alex P. Scpton^{1,2*} & Arvin C. Dar^{1,2}

Deregulation of the Ras–mitogen activated protein kinase (MAPK) pathway is an early event in many different cancers and a key driver of resistance to targeted therapies¹. Sustained signalling through this pathway is caused most often by mutations in K-Ras, which biochemically favours the stabilization of active RAF signalling complexes². Kinase suppressor of Ras (KSR) is a MAPK scaffold^{3–5} that is subject to allosteric regulation through dimerization with RAF^{6,7}. Direct targeting of KSR could have important therapeutic implications for cancer; however, testing this hypothesis has been difficult owing to a lack of small-molecule antagonists of KSR function. Guided by KSR mutations that selectively suppress oncogenic, but not wild-type, Ras signalling, we developed a class of compounds that stabilize a previously unrecognized inactive state of KSR. These compounds, exemplified by APS-2-79, modulate KSR-dependent MAPK signalling by antagonizing RAF heterodimerization as well as the conformational changes required for phosphorylation and activation of KSR-bound MEK (mitogen-activated protein kinase kinase). Furthermore, APS-2-79 increased the potency of several MEK inhibitors specifically within Ras-mutant cell lines by antagonizing release of negative feedback signalling, demonstrating the potential of targeting KSR to improve the efficacy of current MAPK inhibitors. These results reveal conformational switching in KSR as a druggable regulator of oncogenic Ras, and further suggest co-targeting of enzymatic and scaffolding activities within Ras–MAPK signalling complexes as a therapeutic strategy for overcoming Ras-driven cancers.

Ras is the most frequently mutated human oncogene. Yet, despite recent breakthroughs, therapeutic options to target Ras-dependent cancers remain limited¹. Studies conducted in several different model systems support the possibility of Ras-targeted interventions via KSR^{3–5,8–10}. However, due to its status as a pseudokinase and role as a non-catalytic regulator of core signalling enzymes^{11–13}, pharmacological approaches that target KSR have been lacking. This is in contrast to current drug discovery and development efforts that have focused extensively on direct inhibitors of the Ras effector kinases RAF, MEK, and ERK¹⁴.

To explore an alternative form of pharmacological modulation and identify Ras–MAPK antagonists via KSR, we focused on large forward genetic screens conducted in flies and worms that identified mutant Ras-selective suppressor alleles in KSR^{3–5}. The studies in flies alone evaluated approximately 900,000 randomly mutated strains searching for genetic modifiers of a Ras(G12V)-dependent rough-eye phenotype¹⁵. We mapped the suppressor alleles onto the primary sequence of KSR (Extended Data Fig. 1a) and a recently determined X-ray crystal structure of the human KSR2 pseudokinase domain in complex with MEK1 and ATP, and noted a high concentration of suppressor mutations immediately adjacent to the KSR ATP-binding pocket (Fig. 1a). On the basis of this analysis, we hypothesized that the RAF and MEK interaction interfaces in KSR may be uncoupled through ligands that

engage the KSR ATP-binding pocket. Specifically, we speculated that small molecules, which bias KSR towards a state similar to that revealed in the KSR2–MEK1–ATP crystal structure, might function as antagonists of KSR-dependent regulation of RAF and MEK.

To identify active-site-directed ligands of KSR, we screened a collection of 176 structurally diverse kinase inhibitors for direct competition of an activity-based probe (ATP^{biotin}) that specifically labels the ATP-binding pocket of purified KSR2–MEK1 complexes (Fig. 1b, c). From this analysis we identified APS-1-68-2 as a competitor of probe-labelling of KSR2–MEK1. This quinazoline–biphenyl ether compound has previously been described as both a Src and epidermal growth factor receptor (EGFR) family kinase inhibitor. Synthetic tailoring of APS-1-68-2 generated highly informative structure–activity relationships (Fig. 1d). For example, deletion of the terminal phenyl group (APS-1-82-1) or extension of the ether linker (APS-2-12) diminished KSR2–MEK1 probe competition. Notably, addition of a single methyl group at the internal phenyl generated a potent probe compound (APS-2-79; IC₅₀ of KSR2 = 120 ± 23 nM), whereas the similar dimethyl substituted compound (APS-3-77) was essentially inactive (IC₅₀ of KSR2 > 10,000 nM).

To assess the biological function of these compounds as Ras–MAPK pathway antagonists, we developed a simplified cell-based reconstitution system to directly monitor KSR-driven MAPK signalling (Fig. 1e). This system, in which cellular MAPK signalling is dependent on KSR expression, was found to be sensitive to known Ras suppressor mutations in KSR (Fig. 1f). Likewise, APS-2-79 also suppressed KSR-stimulated MEK and ERK phosphorylation (Fig. 1g; *P* < 0.005 lanes 1 versus 2). The suppression of MAPK signalling by APS-2-79 was dependent on direct targeting of KSR as an active site mutant (KSR(A690F)), which has previously been demonstrated to stimulate KSR-based MAPK outputs independent of ATP-binding¹⁶, significantly diminished the activity of APS-2-79 (Fig. 1g; lanes 5 versus 6, NS; lanes 2 versus 6, *P* < 0.005). Notably, the negative control for KSR-binding (analogue APS-3-77; see Extended Data Fig. 2b, c for comparative selectivity profiling) was inactive, whereas a positive-control RAF inhibitor, dabrafenib, was active irrespective of the KSR-mutational status (Fig. 1g). Therefore, on the basis of similarity in phenotype and also direct-binding activity, we identify APS-2-79 as a small-molecule mimic of KSR alleles that suppress oncogenic Ras mutations.

KSR-based activity of APS-2-79 as a MAPK antagonist was further evaluated using reconstitution assays. Dose-dependent phosphorylation of MEK on Ser218/Ser222 by RAF *in vitro* could be enhanced at least fivefold in the presence of KSR (Extended Data Fig. 3a–c). KSR-stimulated MEK phosphorylation by RAF was markedly reduced by the addition of APS-2-79, but not by APS-3-77 (Extended Data Fig. 3d, e). APS-2-79 was inactive when KSR was absent or when the KSR2(A690F) mutant was used for *in vitro* assays (Extended Data Fig. 3d, f, g), suggesting that the activity of APS-2-79 derives from direct targeting of KSR. Indeed, APS-2-79 lacked direct activity against the highly homologous active RAF family kinases, including recombinant

¹Department of Oncological Sciences, The Tisch Cancer Institute, The Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. ²Department of Structural and Chemical Biology, The Tisch Cancer Institute, The Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA.

*These authors contributed equally to this work.

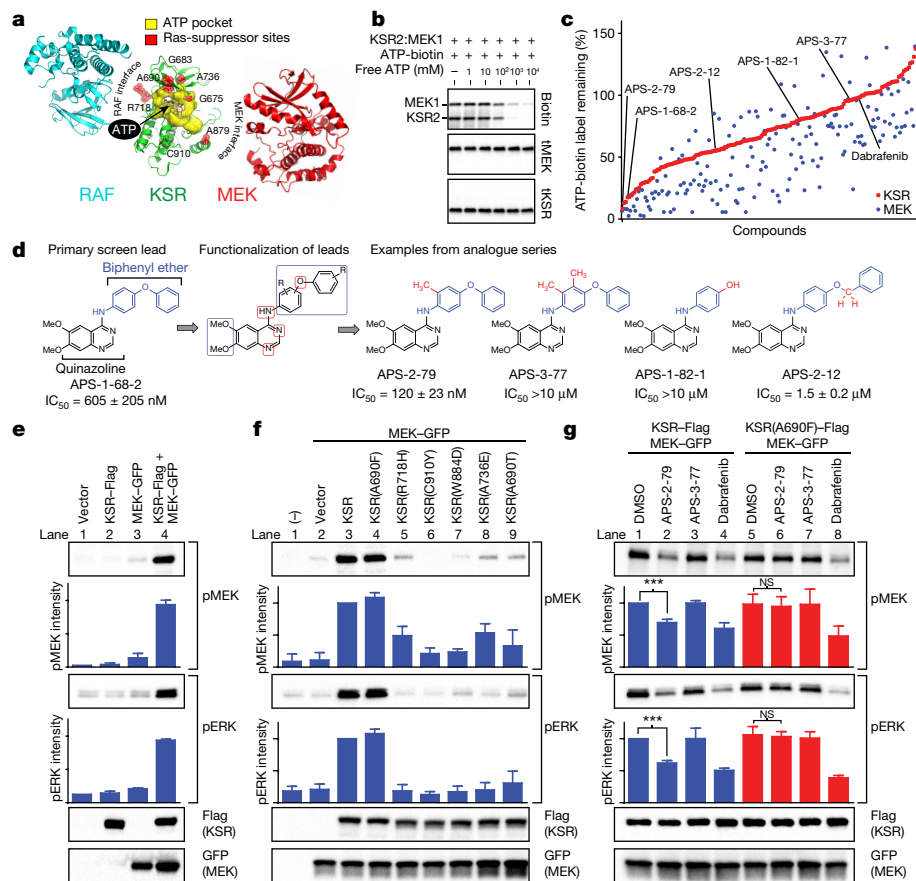


Figure 1 | The small molecule APS-2-79 mimics KSR alleles that suppress oncogenic Ras mutations. a, Oncogenic Ras-suppressor mutations (red) localize to the ATP-binding pocket (yellow), as well as RAF- and MEK- interaction interfaces, in KSR. Shown is the putative structure of the RAF–KSR–MEK complex⁷. **b**, An activity-based probe (ATP^{biotin}) specifically labels the ATP-binding pockets of purified KSR2–MEK1 complexes. 2 μ M of ATP^{biotin} was incubated with KSR2–MEK1 in the presence of the indicated concentrations of free ATP. Biotin, total MEK, and total KSR western blots are shown. **c**, A kinase inhibitor screen for direct competitors of probe-labelling in purified KSR2–MEK1 complexes provides informative structure–activity relationships data. **d**, Chemical structures of leads. IC_{50} values (mean \pm s.d.; $n = 2$ biological replicates) against ATP^{biotin} probe-labelling of KSR2 are listed below structures. **e**, Co-expression of full-length KSR–Flag and MEK1–GFP

BRAF and CRAF, or cellular BRAF(V600E) (Extended Data Figs 2, 3, 4a). Therefore, on the basis of reconstitution and selectivity assays, we conclude that APS-2-79 functions as an antagonist of MEK phosphorylation by RAF through direct binding of the KSR active site.

Notably, we found that a previously described ATP-competitive and active-state binder of KSR termed ASC24 (ref. 7), in contrast to APS-2-79, did not antagonize KSR-dependent MEK phosphorylation by RAF (Extended Data Fig. 3d, e), suggesting that inhibition of catalytic activity alone in KSR is insufficient to block MAPK signalling. Consistent with this notion, removal of putative KSR phosphorylation sites⁷ in MEK neither impeded MAPK signalling nor blocked the inhibitory activity of APS-2-79 within cells (Extended Data Fig. 4b).

Previous studies established that genetic suppressors in KSR may impede RAF-induced conformational changes in KSR required for MEK activation or destabilize KSR–MEK and KSR–RAF complexes^{6,7,12,17–19}. To distinguish between such possible modes of action, we determined an X-ray crystal structure of the KSR2–MEK1 complex bound to APS-2-79 (Fig. 2a). In the APS-2-79-bound state, KSR2 binds MEK1 in a 1:1 fashion within a quaternary arrangement that is nearly identical to the ATP-bound state of KSR2–MEK1 complexes⁷ (Extended Data Fig. 5). Within both states, KSR2 and

leads to enhanced MAPK signalling within 293H cells, as visualized by immunoblotting for phosphorylated MEK and ERK. **f**, MAPK activation is sensitive to known genetic suppressor mutations in KSR. A690F is a KSR mutant predicted to signal independent of ATP-binding¹⁶. W884D is a loss-of-function mutation predicted based on structural analysis. Note, human KSR2 numbering used here and throughout. **g**, APS-2-79 impedes KSR-stimulated MAPK signalling within cells by wild-type KSR but not a control mutant (KSR(A690F)). Cells were treated with 5 μ M of APS-2-79, APS-3-77, or dabrafenib for 2 h. In **e–g**, cells were collected for western blot analysis 24 h after transfection. Error bars indicate the mean \pm s.d. ($n = 3$ biological replicates). Signals were normalized relative to lane 1 (**e** and **g**) or 3 (**f**). NS, not significant. *** $P < 0.0005$ by two-tailed unpaired t -testing.

MEK1 bind via a face-to-face arrangement mediated largely through reciprocal helix α G and activation segment interactions, and KSR2 homodimerizes through the N-lobe along a crystallographic two-fold symmetry axis producing a hetero-tetramer of KSR2–MEK1 dimers.

In the APS-2-79-bound state, only KSR2 was found to possess strong electron density that could be assigned to APS-2-79 (Extended Data Fig. 6a, b). Two portions of APS-2-79 engage distinct regions in KSR2. First, the biphenyl ether extends to a sub-pocket within KSR2, defined by Thr739, Arg692, Asp803 and a hydrophobic shell composed of Phe725, Tyr714 and Phe804 (Fig. 2b, c). Stacking interactions between the terminal phenyl in APS-2-79, and Phe725, Tyr714 and Phe804 in particular are expected to provide strong interactions between KSR2 and APS-2-79 through the arrangement of a four-member aromatic-pair network (Fig. 2b, c). The existence of this network was substantiated by removal of the terminal phenyl in APS-2-79-like compounds, which greatly diminished competition of ATP^{biotin} probe-labelling in KSR2 (Extended Data Fig. 7; APS-1-68-2 versus APS-1-70-1 and APS-1-82-1). This network of aromatic-pair interactions, in addition to other amino acid substitutions, probably contributes to the selectivity of APS-2-79 for KSR over RAF (Extended Data Fig. 6c, e). Second, a hydrogen bond between the N1 in the quinazoline core of APS-2-79 and the

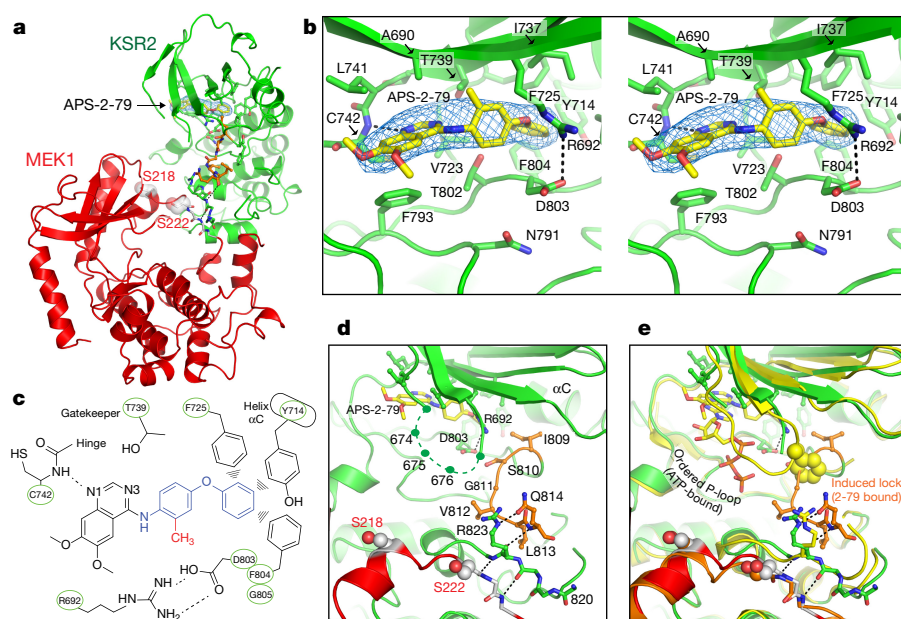


Figure 2 | Structural analysis of APS-2-79 bound to the KSR2-MEK1 complex. **a**, The KSR2-MEK1-APS-2-79 complex. Highlighted are two key phospho-regulatory residues in MEK1, Ser218 and Ser222. **b**, Magnified stereo view of interactions between KSR2 and APS-2-79. $F_o - F_c$ omit map contoured at 3.5σ , generated with APS-2-79 omitted, is represented as a blue mesh. **c**, Schematic of the APS-2-79 binding site within KSR2. **d**, Magnified view of the KSR2 active site bound to APS-2-79, including the 'induced lock' (residues I809-Q814; orange). The disordered P-loop is highlighted by a dashed line. **e**, Overlay between the ATP-bound (yellow) and APS-2-79-bound states of KSR2.

backbone at Cys742 further mediates APS-2-79-KSR2 interactions. Notably, functionalization of the N1 with a methyl group (APS-3-6) greatly diminished KSR2-MEK1 activity, whereas replacement of the N3 with -CH (APS-2-16) was moderately tolerated (Extended Data Fig. 7). Therefore, on the basis of crystallographic analysis and also structure-activity relationships data from our analogue series, APS-2-79 binds directly to KSR2 within the KSR2-MEK1 complex.

In both the APS-2-79- and ATP-bound states of KSR2-MEK1, KSR2 directly engages the activation segment of MEK1, burying the Ser218-Ser222 region and presumably shielding this segment of MEK from promiscuous phosphorylation. The KSR2-MEK1-APS-2-79 structure revealed a portion of KSR2 that was not previously modelled in the ATP-bound complex (Extended Data Fig. 6d). This region, encompassing residues Ile809 to Gln814, which we refer to as the induced lock, forms an extension of the activation segment C terminus to the conserved DFG motif, and forms an anti-parallel β -strand with the peptide sequence centred around Arg823 in KSR2 (Fig. 2d). Additionally, the ordering of residues Ile809 to Gln814 in KSR2 occurs at the expense of

disorder of residues 674 to 676 in the P-loop, which in the ATP-bound state directly coordinates the β and γ phosphates (Fig. 2e). The two modes by which ATP and APS-2-79 affect KSR2-based interactions on MEK appear mutually exclusive as both ligands induce conformations that would otherwise clash with one another (Fig. 2e). We interpret these structures to suggest that APS-2-79 stabilizes an inactive state of KSR2 characterized by reinforcement of negative regulatory interactions. Indeed, APS-2-79 behaves as a KSR-dependent antagonist of RAF-mediated MEK phosphorylation by shifting the equilibrium of KSR-MEK complexes so as to populate the OFF state (Extended Data Fig. 5c).

Comparison of the ATP-bound and APS-2-79-bound states of KSR2-MEK1 suggested that APS-2-79 antagonizes RAF phosphorylation on MEK indirectly by impeding KSR-RAF heterodimers. As well as APS-2-79 binding, the dimer interface of KSR2, including residues Trp685 and His686, demonstrated perturbations relative to the ATP-bound conformation (Fig. 3a, Extended Data Fig. 8). To investigate directly the effect of APS-2-79 on KSR2-BRAF dimerization, we used bio-layer

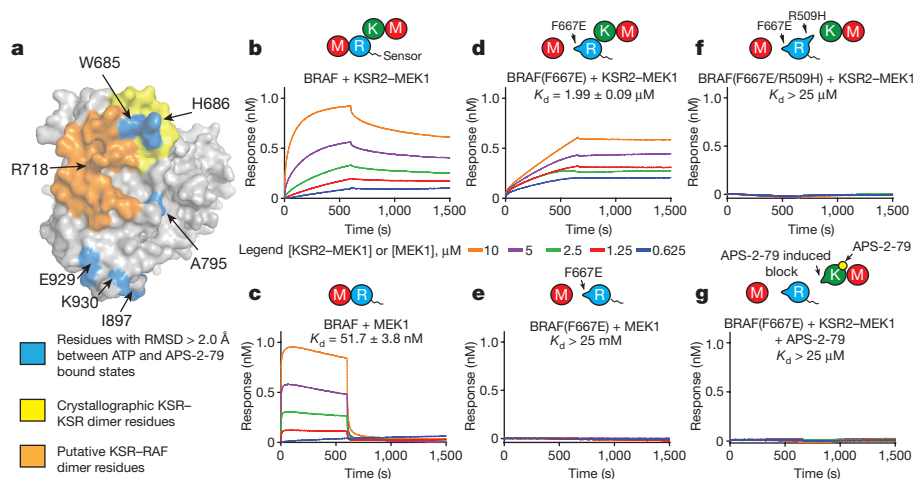


Figure 3 | APS-2-79 impedes higher order assembly of RAF-KSR-MEK complexes. **a**, Mapping of residues with a root-mean-square (r.m.s.) deviation of >2.0 Å between the ATP- and APS-2-79-bound states of KSR2-MEK1 (blue) highlights alterations at contact residues Trp685 and His686 within the putative KSR-RAF heterodimer interface. **b-g**, BRAF and BRAF mutants (F667E and/or R509H) were immobilized on sensor-heads

and KSR2-MEK1 or MEK1 assembly was monitored using bio-layer interferometry. Association occurred from 0 to 660 s and dissociation was monitored thereafter up to 1500 s. APS-2-79 was added in the presence of KSR2-MEK1 at a concentration of $25 \mu\text{M}$. K_d values represent the mean \pm s.e.m. derived from global fitting of all 5 binding curves.

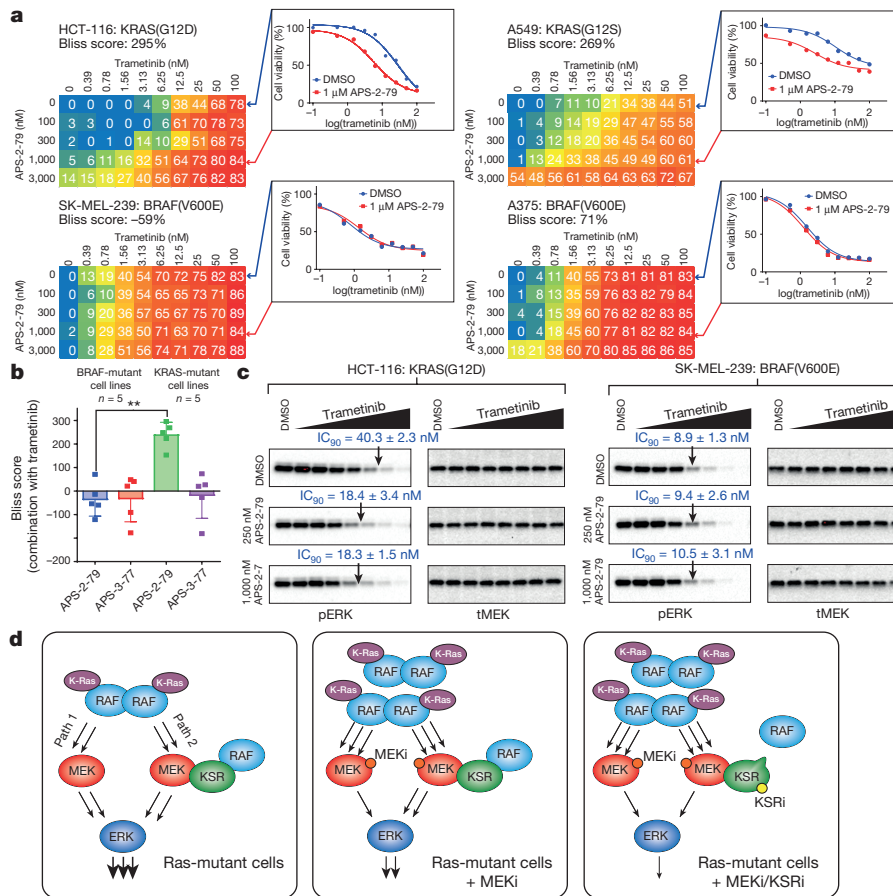


Figure 4 | APS-2-79 Enhances the efficacy of the clinical MEK inhibitor trametinib within cancer cell lines containing K-Ras mutations.

a, Dose-responses of APS-2-79 and trametinib (MEKi) on viability of K-Ras-mutant (HCT-116, A549) and BRAF-mutant (A375, SK-MEL-239) cell lines. Bliss scores represent the mean calculated from two biological replicates of the depicted concentration matrices. Numbers listed within synergy matrices, which represent the percentage of growth inhibition relative to DMSO controls, are the mean of the replicates. Insets highlight dose-responses of trametinib in the absence or presence of 1 μ M APS-2-79 (points along each line represent the mean of two biological replicates). **b**, Synergy, as determined by Bliss independence scores (see Methods), for combinations of APS-2-79 and APS-3-77 with trametinib. Error bars represent the mean \pm s.d. of Bliss scores as determined in **a** and Extended

Data Fig. 9, derived from either K-Ras-mutant and BRAF-mutant cancer cell lines ($n = 5$ for each). $**P < 0.005$ by two-tailed unpaired t -testing. **c**, Pathway analysis suggests that the increased potency of trametinib in the presence of APS-2-79 occurs through enhanced downregulation of Ras-MAPK signalling (as measured by phospho-ERK). HCT-116 and SK-MEL-239 cells were treated for 48 h with increasing concentrations of trametinib combined with DMSO, 250 nM, and 1 μ M APS-2-79. IC₉₀ values represent the mean \pm s.d. ($n = 2$ biological replicates). **d**, Model for synergy between the MEK inhibitor (MEKi) trametinib and the KSR inactive state binder (KSRi) APS-2-79. APS-2-79 enhances the efficacy of trametinib by antagonizing MEKi-induced Ras-MAPK signalling complexes.

inferometry (BLI) to monitor real-time association and dissociation of KSR2-MEK1 or free MEK1 to a sensor tethered with immobilized BRAF. In control experiments, we found that KSR2-MEK1 complexes did not associate with immobilized BRAF in a 1:1 fashion (Fig. 3b), probably owing to the formation of higher order BRAF-KSR2-MEK1 complexes. In contrast, BRAF bound to free MEK1 in a 1:1 fashion with a dissociation constant (K_d) = 51 ± 3.8 nM (Fig. 3c), which is in close agreement to published work²⁰.

To specifically monitor KSR2-BRAF dimerization relative to other possible interactions, we identified a mutation in BRAF(F667E) that eliminates binding to free MEK but not KSR2-MEK1 complexes (Fig. 3d, e). KSR2-MEK1 interacted in a 1:1 fashion with the BRAF(F667E) mutant with a K_d of 1.99 ± 0.09 μ M; closely matching previously published BRAF-BRAF dimerization values⁶. Notably, the addition of a secondary mutation, known to perturb KSR2-BRAF dimers (BRAF(F667E/R509H); Fig. 3f), completely abrogated any binding signal between KSR2-MEK1 and BRAF. In the presence of APS-2-79, the KSR2-BRAF(F667E) dimers did not associate (Fig. 3g), consistent with the prediction of the crystal structure suggesting that APS-2-79 may impede RAF-KSR dimers. In contrast, the control compound APS-3-77 did not impede KSR2-BRAF interactions (Extended

Data Fig. 8c). Therefore, we conclude that BRAF can dimerize with KSR2-MEK1 complexes directly via KSR2, and this interaction is antagonized by APS-2-79.

Ras mutations occur in approximately 25% of all cancer patients and are highly associated with poor response to therapy¹. Significant progress has been made in targeting BRAF(V600E)-mutant melanoma, however RAF and MEK inhibitors have failed to achieve significant clinical efficacy in Ras-mutant disease owing in part to mechanisms of inhibitor-induced transactivation and feedback, respectively²¹. MEK-inhibitor feedback has been characterized by upstream Ras activation and induction of higher-order RAF-RAF and also RAF-KSR complexes²²⁻²⁴. In an engineered cell system, we found that a Ras-suppressor allele (R718H^{6,7}) within KSR reduced MEK inhibitor-induced feedback (Extended Data Fig. 4c), suggesting the possibility that KSR heterodimerization may limit the efficacy of MEK inhibitors. Owing to the more pronounced role of KSR in Ras-mutant, as opposed to RAF-mutant signalling^{15,25}, and the ability of APS-2-79 to impede KSR-RAF heterodimerization, we hypothesized that stabilization of the KSR-inactive state (KSRi) via APS-2-79 may potentiate the effect of MEK inhibitors by limiting feedback in Ras-mutant models. We therefore tested for synergy of APS-2-79 with MEK inhibitors in Ras-mutant cell lines, and used RAF-mutant cell lines as controls.

We found that APS-2-79 shifted the cell viability dose response to trametinib in Ras-mutant cell lines HCT-116 and A549, but not BRAF mutant cell lines SK-MEL-239 and A375 (Fig. 4a). Although the cellular effects of APS-2-79 alone were modest, combination analysis over full concentration matrices revealed that KSRi synergizes with trametinib, and other MEK inhibitors (Extended Data Fig. 9a), specifically in KRAS mutant cell lines (Fig. 4b). APS-3-77, and additional control compounds (Extended Data Fig. 9b and 10), did not demonstrate Ras-mutant-specific synergy, supporting the hypothesis that the enhanced activity of trametinib when combined with APS-2-79 depends on co-modulation of KSR. To determine the possible mechanism for APS-2-79 and trametinib synergy, we examined MAPK signalling and found that APS-2-79 treatment caused a twofold enhancement in the IC₅₀ of trametinib on ERK phosphorylation in the Ras-mutant HCT-116 cell line but not the RAF-mutant SK-MEL-239 cell line (Fig. 4c, Extended Data Fig. 4d). The data presented here provide proof-of-concept for the use of KSRi to overcome a key liability of a clinical MEK inhibitor in K-Ras mutant cells. Indeed, we posit stabilization of the KSRi as a mechanism to impede feedback activated Ras–MAPK signalling induced by MEK inhibition (Fig. 4d).

Here we have identified a unique conformation in KSR through the discovery of APS-2-79. This compound offers a foundation for the development of a new class of targeted therapies based on stabilization of the KSR inactive state. Future efforts will be directed towards improving the pharmacological properties of APS-2-79 to enable *in vivo* and clinical studies. In general, the stabilization of conformational states with small-molecule modulators may be an effective strategy to target other pseudokinases^{26,27}. Furthermore, the results presented here, using KSRi in combination with clinical MEK inhibitors, suggests a mechanism to improve the efficacy of inhibitors that target enzymatically active kinases through co-modulation of pseudokinase–active kinase signalling complexes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 November 2015; accepted 22 July 2016.

Published online 24 August 2016.

- Stephen, A. G., Esposito, D., Bagni, R. K. & McCormick, F. Dragging ras back in the ring. *Cancer Cell* **25**, 272–281 (2014).
- Lavoie, H. & Therrien, M. Regulation of RAF protein kinases in ERK signalling. *Nat. Rev. Mol. Cell Biol.* **16**, 281–298 (2015).
- Kornfeld, K., Hom, D. B. & Horvitz, H. R. The *ksr-1* gene encodes a novel protein kinase involved in Ras-mediated signaling in *C. elegans*. *Cell* **83**, 903–913 (1995).
- Sundaram, M. & Han, M. The *C. elegans ksr-1* gene encodes a novel Raf-related kinase involved in Ras-mediated signal transduction. *Cell* **83**, 889–901 (1995).
- Therrien, M. *et al.* KSR, a novel protein kinase required for RAS signal transduction. *Cell* **83**, 879–888 (1995).
- Rajakulendran, T., Sahmi, M., Lefrançois, M., Sicheri, F. & Therrien, M. A dimerization-dependent mechanism drives RAF catalytic activation. *Nature* **461**, 542–545 (2009).
- Brennan, D. F. *et al.* A Raf-induced allosteric transition of KSR stimulates phosphorylation of MEK. *Nature* **472**, 366–369 (2011).
- Fernandez, M. R., Henry, M. D. & Lewis, R. E. Kinase suppressor of Ras 2 (KSR2) regulates tumor cell transformation via AMPK. *Mol. Cell Biol.* **32**, 3718–3731 (2012).
- Le Borgne, M., Filbert, E. L. & Shaw, A. S. Kinase suppressor of Ras 1 is not required for the generation of regulatory and memory T cells. *PLoS One* **8**, e57137 (2013).
- Nguyen, A. *et al.* Kinase suppressor of Ras (KSR) is a scaffold which facilitates mitogen-activated protein kinase activation *in vivo*. *Mol. Cell Biol.* **22**, 3035–3045 (2002).
- Michaud, N. R. *et al.* KSR stimulates Raf-1 activity in a kinase-independent manner. *Proc. Natl Acad. Sci. USA* **94**, 12792–12796 (1997).
- Roy, F., Laberge, G., Douziech, M., Ferland-McCollough, D. & Therrien, M. KSR is a scaffold required for activation of the ERK/MAPK module. *Genes Dev.* **16**, 427–438 (2002).
- Stewart, S. *et al.* Kinase suppressor of Ras forms a multiprotein signaling complex and modulates MEK localization. *Mol. Cell Biol.* **19**, 5523–5534 (1999).
- Samatar, A. A. & Poulikakos, P. I. Targeting RAS-ERK signalling in cancer: promises and challenges. *Nat. Rev. Drug Discov.* **13**, 928–942 (2014).
- Karim, F. D. *et al.* A screen for genes that function downstream of Ras1 during *Drosophila* eye development. *Genetics* **143**, 315–329 (1996).
- Hu, J. *et al.* Allosteric activation of functionally asymmetric RAF kinase dimers. *Cell* **154**, 1036–1046 (2013).
- Lavoie, H. *et al.* Inhibitors that stabilize a closed RAF kinase domain conformation induce dimerization. *Nat. Chem. Biol.* **9**, 428–436 (2013).
- Douziech, M., Sahmi, M., Laberge, G. & Therrien, M. A KSR/CNK complex mediated by HYP, a novel SAM domain-containing protein, regulates RAS-dependent RAF activation in *Drosophila*. *Genes Dev.* **20**, 807–819 (2006).
- McKay, M. M., Ritt, D. A. & Morrison, D. K. Signaling dynamics of the KSR1 scaffold complex. *Proc. Natl Acad. Sci. USA* **106**, 11022–11027 (2009).
- Haling, J. R. *et al.* Structure of the BRAF-MEK complex reveals a kinase activity independent role for BRAF in MAPK signaling. *Cancer Cell* **26**, 402–413 (2014).
- Chapman, P. B., Solit, D. B. & Rosen, N. Combination of RAF and MEK inhibition for the treatment of BRAF-mutated melanoma: feedback is not encouraged. *Cancer Cell* **26**, 603–604 (2014).
- Hatzivassiliou, G. *et al.* Mechanism of MEK inhibition determines efficacy in mutant KRAS- versus BRAF-driven cancers. *Nature* **501**, 232–236 (2013).
- Lito, P. *et al.* Disruption of CRAF-mediated MEK activation is required for effective MEK inhibition in KRAS mutant tumors. *Cancer Cell* **25**, 697–710 (2014).
- Sos, M. L. *et al.* Oncogene mimicry as a mechanism of primary resistance to BRAF inhibitors. *Cell Reports* **8**, 1037–1048 (2014).
- McKay, M. M., Ritt, D. A. & Morrison, D. K. RAF inhibitor-induced KSR1/B-RAF binding and its effects on ERK cascade signaling. *Curr. Biol.* **21**, 563–568 (2011).
- Xie, T. *et al.* Pharmacological targeting of the pseudokinase Her3. *Nat. Chem. Biol.* **10**, 1006–1012 (2014).
- Zeqiraj, E., Filippi, B. M., Deak, M., Alessi, D. R. & van Aalten, D. M. Structure of the LKB1-STRAD-MO25 complex reveals an allosteric mechanism of kinase activation. *Science* **326**, 1707–1711 (2009).

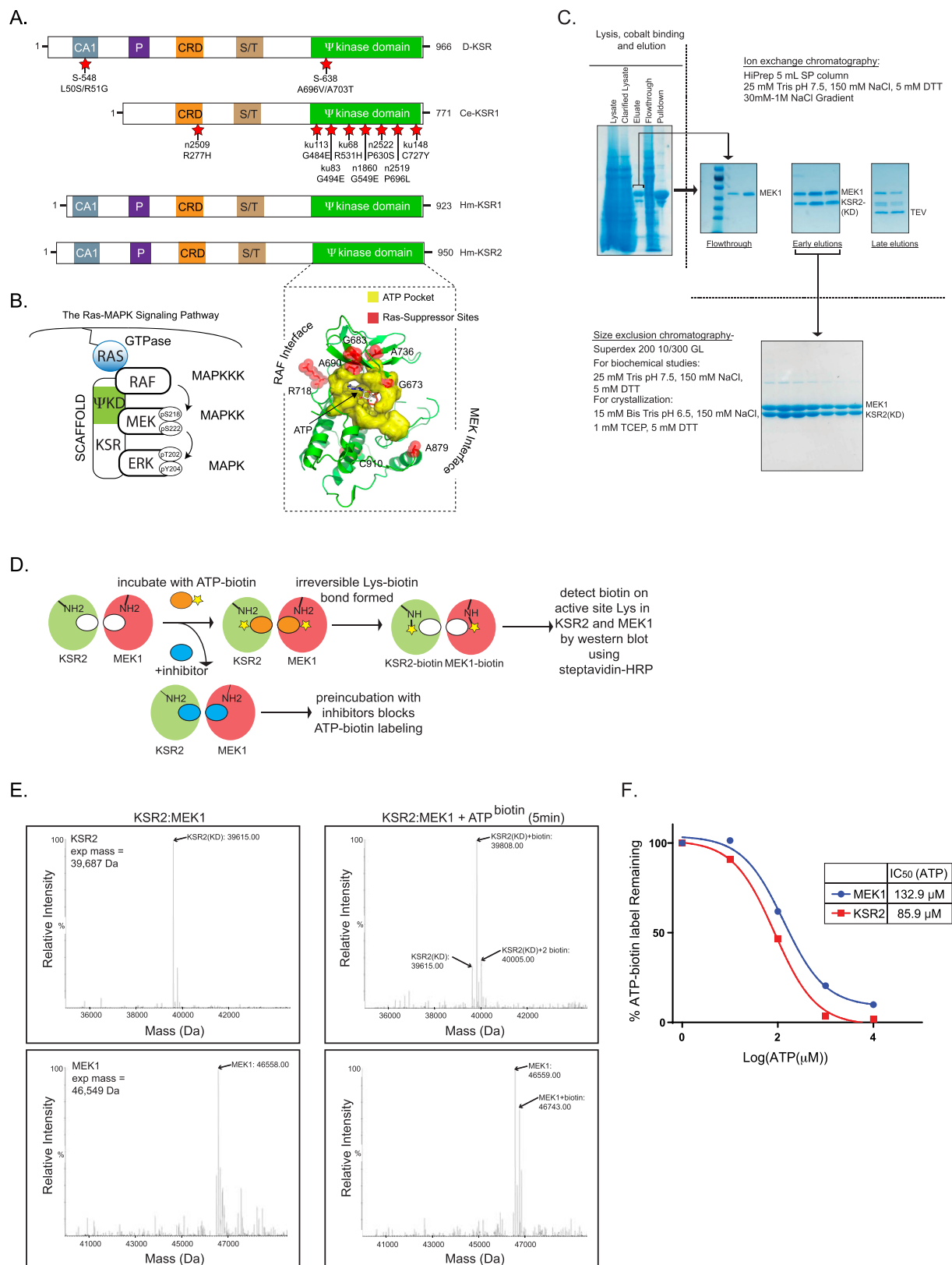
Supplementary Information is available in the online version of the paper.

Acknowledgements We thank K. Shokat, R. Fisher, R. Cagan, S. Aaronson, M. Lazarus, E. Bernstein, and members of the Dar laboratory for comments on the manuscript; A. Maldonado and L. Silber for technical support; and staff at the Advanced Photon Source (Beamline 23-ID-B) for help with X-ray diffraction experiments. The Dar laboratory is supported by innovation awards from the NIH (1DP2CA186570-01) and Damon Runyon-Rachleff Foundation. A.C.D. is a Pew-Stewart Scholar in Cancer Research and Young Investigator of the Pershing-Square Sohn Cancer Research Alliance.

Author Contributions N.S.D. conducted biochemical, structural, and cell-line studies. A.P.S. synthesized compounds. A.C.D. supervised research. All authors analysed data.

Author Information Coordinates and structure factors have been deposited with the Protein Data Bank under accession code 5KRR. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.C.D. (arvin.dar@mssm.edu).

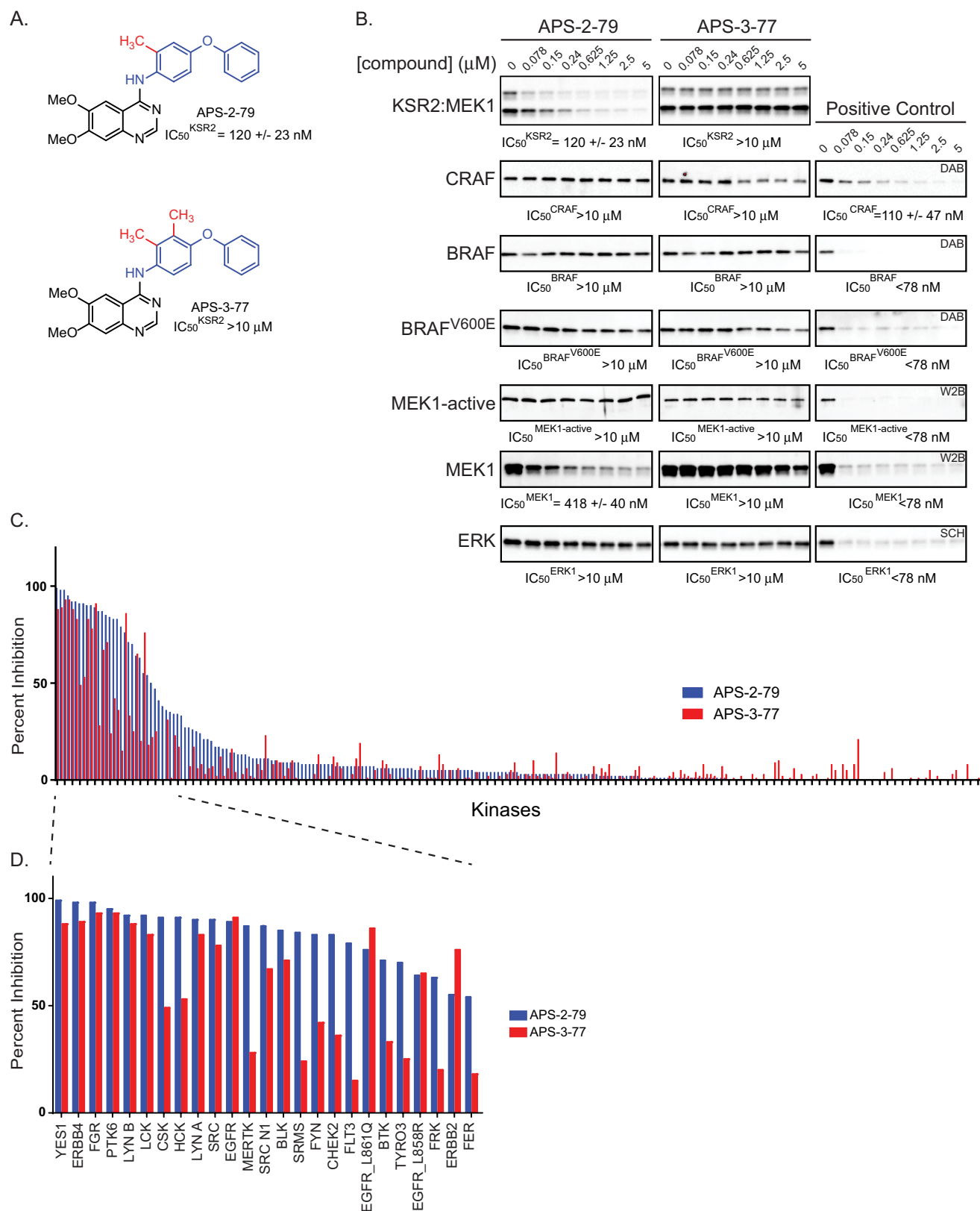
Reviewer Information *Nature* thanks G. Bollag and the other anonymous reviewer(s) for their contribution to the peer review of this work.



Extended Data Figure 1 | See next page for caption.

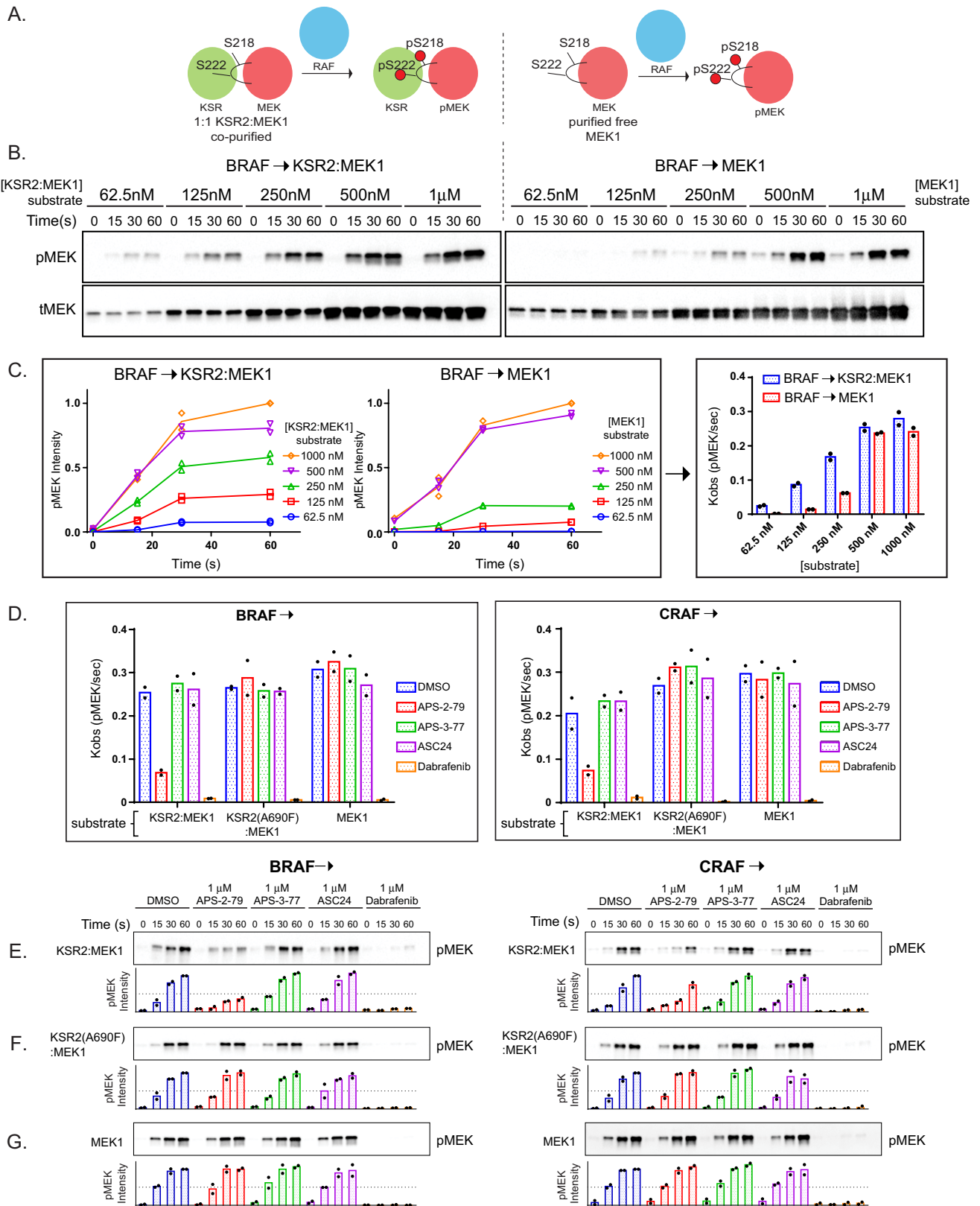
Extended Data Figure 1 | Projection of Ras(G12V) suppressor alleles onto the primary and tertiary structure of KSR. **a**, Schematic representation of KSR from *Drosophila*, *Caenorhabditis elegans*, and KSR1 or KSR2 from humans. Suppressor mutations within KSR identified from forward genetic screens are highlighted with red stars. Allele names and corresponding mutations are given^{3–5,15}. Two alleles in KSR found in the *Drosophila* screen are shown; one encoding for substitutions in a coil-coil SAM domain (CC-SAM) at the N terminus of *Drosophila* KSR (S548) and a second mutant in the predicted ATP-binding pocket of the KSR pseudokinase domain (S638). Eight distinct alleles were described in two separate studies conducted in *C. elegans*. The vast majority of the mutants localize to the pseudokinase domain of KSR and in particular ATP-contact residues (yellow). Residues highlighted in red and shown in the lower panel correspond to the human KSR2 residue equivalents of suppressor mutations found in *Drosophila* and *C. elegans* orthologues. **b**, KSR is a scaffold for the Ras–MAPK signalling pathway. Phosphorylation of MEK1/2 at Ser218 and Ser222 by RAF, or ERK1/2 via phosphorylation at Thr202 and Tyr204 by MEK, are key events in signalling through

the Ras–MAPK signalling pathway. **c**, Purification of the KSR2–MEK1 complex from insect cells. The KSR2 pseudokinase domain (KSR2(KD)) and MEK1 were co-expressed using the SF21 insect cell system. Lysis was performed by one freeze–thaw and sonication. Lysates were incubated with cobalt resin for 2 h and KSR2(KD)–MEK1 was eluted using a high-imidazole buffer. Eluate was then incubated with tomato etch virus (TEV) protease and λ -phosphatase overnight. The mixture was then applied to an ion-exchange column (Sp-HP) to separate stoichiometric KSR2–MEK1 complexes from free MEK1 and TEV. Fractions containing KSR2–MEK1 were applied to a gel-filtration column for final purification. **d**, Schematic of the ATP^{biotin} probe-labelling assay on KSR2–MEK1 complexes and screen for inhibitors. **e**, ATP^{biotin} directly labels KSR2 and MEK1 within purified complexes. Deconvoluted mass spectrum for KSR2–MEK1 complexes incubated with ATP^{biotin}. KSR2 and MEK1 spectra are included in the top and bottom panels, respectively. **f**, Graphical representation for ATP^{biotin} probe-labelling of KSR2–MEK1 complexes in the presence of increasing free ATP as shown in Fig. 1b. Corresponding IC₅₀ values listed for both KSR2 and MEK1.



Extended Data Figure 2 | APS-2-79 and APS-3-77 are positive and negative binders of KSR2. **a**, Chemical structures of APS-2-79 and APS-3-77 with respective IC_{50} values (mean \pm s.d.; $n = 2$ biological replicates) for KSR2. **b**, Representative western blot images of *in vitro* ATP^{biotin} competition assays using recombinant MAPK family member proteins. Probe-labelling of the indicated kinases were measured in the presence of increasing concentrations of APS-2-79, APS-3-77, or a positive control compound. For CRAF, BRAF, and BRAF(V600E), the positive control was dabrafenib; for MEK1, the ATP-competitive inhibitor termed Wyeth-2b

(ref. 28); and for ERK, SCH722984 (ref. 29). The listed IC_{50} values include mean \pm s.d. based on two biological replicates. **c**, APS-3-77 and APS-2-79 share partially overlapping kinome-wide inhibitory profiles. The graph shows the percentage of inhibition of APS-2-79 and APS-3-77 (both at $1 \mu\text{M}$) against 246 kinases. The raw data for this graph is in Supplementary Table 1. **d**, Inset showing the 25 kinases most inhibited by APS-2-79 and APS-3-77. Kinases with near-equal sensitivity to these inhibitors as measured here include YES1, ERBB4, and FGR; variable sensitivity kinases include CSK, HCK, and MERTK.

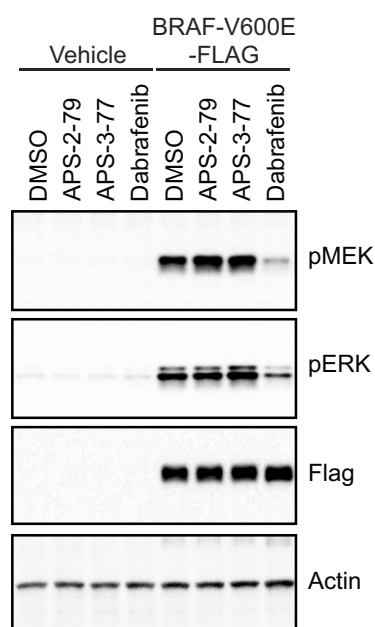


Extended Data Figure 3 | See next page for caption.

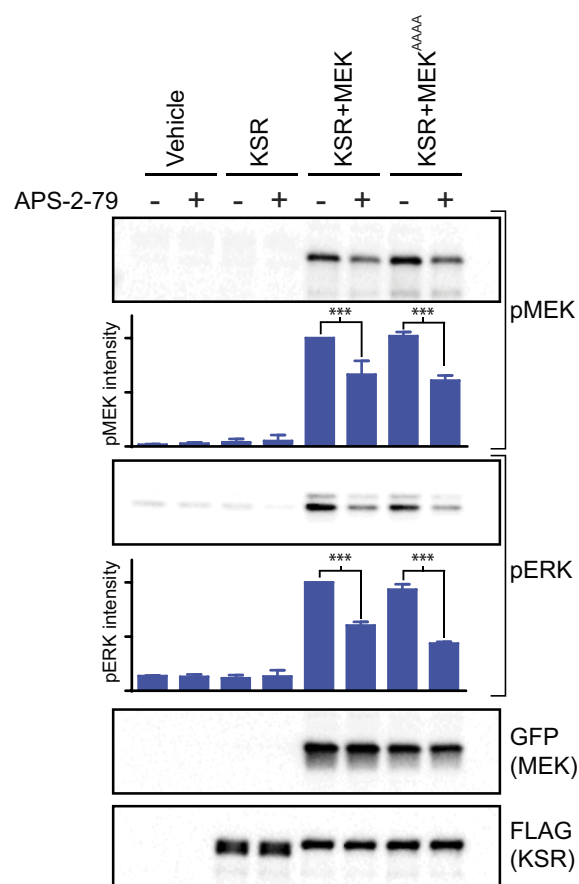
Extended Data Figure 3 | APS-2-79 hinders RAF-mediated MEK phosphorylation in a KSR-dependent manner. **a**, Schematic of the RAF phosphorylation assay of free KSR2–MEK1 and MEK1. **b**, Phosphorylation of the indicated concentrations of MEK1 and the KSR2–MEK1 complex by BRAF (200 nM) in the presence of 1 mM ATP. Representative blots for phospho-MEK (top; as detected using a MEK1/2(pS218/pS222) antibody) and total MEK (tMEK; bottom) are shown. **c**, Plots of pMEK versus time (seconds) at various concentrations of MEK1 and the KSR2–MEK1 complex. Bands were quantified and the phospho-MEK signal normalized relative to lane 20 in both panels. Data points of two biological replicates are included along each line. The rate of MEK phosphorylation (Kobs; pMEK per second; far right) are represented in bar graphs and are derived from the linear phase of the plots in the

left hand panels. Bars represent mean of two biological replicates; values for each replicate are shown as points. **d**, Rates of BRAF (left) and CRAF (right) phosphorylation of the indicated MEK complexes (KSR2–MEK1; KSR2(A690F)–MEK1; and free MEK). Bars represent mean of two biological replicates; values for each replicate are shown as points. **e–g**, APS-2-79 inhibits BRAF and CRAF phosphorylation of MEK in a KSR-dependent manner. Phosphorylation of 500-nM KSR2–MEK1 (**e**), or KSR2(A690F)–MEK1 (**f**), and MEK1 (**g**) by BRAF (200 nM) or CRAF (10 nM) in the presence of 1-mM ATP and the indicated inhibitors. Representative western blots of phospho-MEK (as detected using a MEK1/2pS218/pS222 antibody) are shown. Bars represent mean of two biological replicates; individual data points of each replicate are shown.

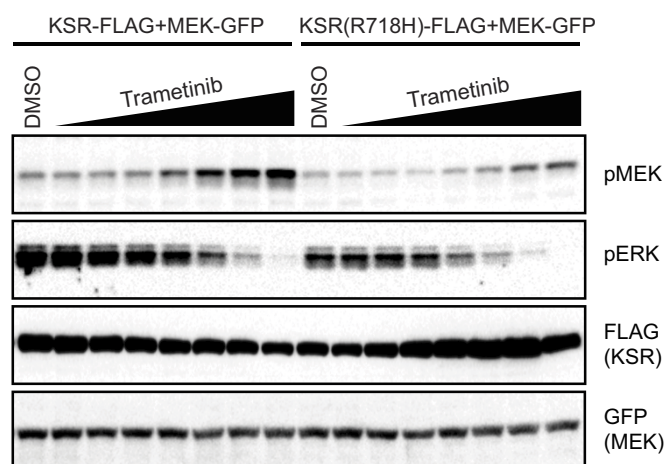
A.



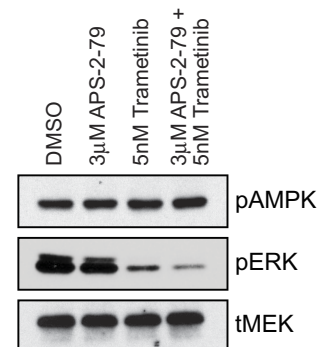
B.



C.



D.

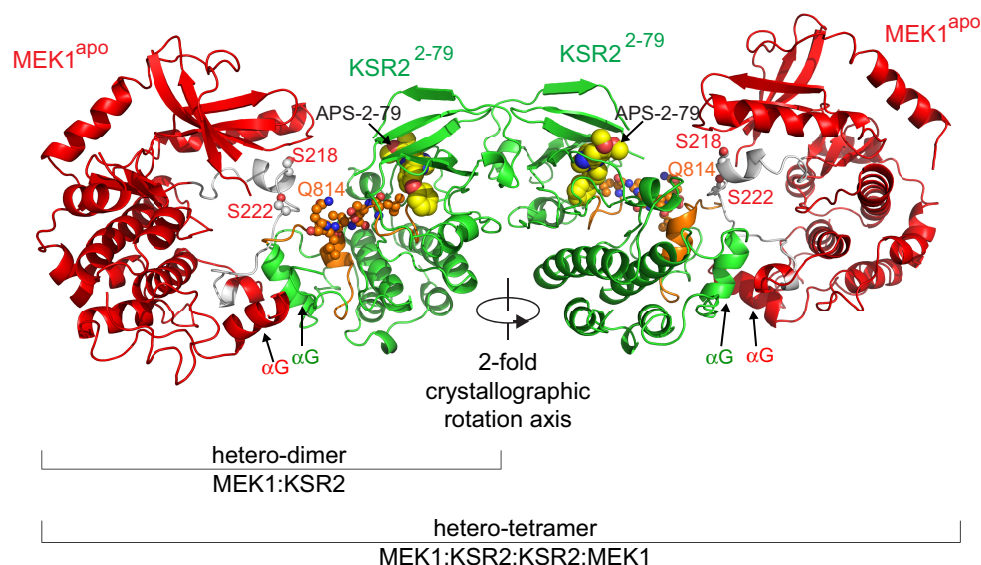


Extended Data Figure 4 | See next page for caption.

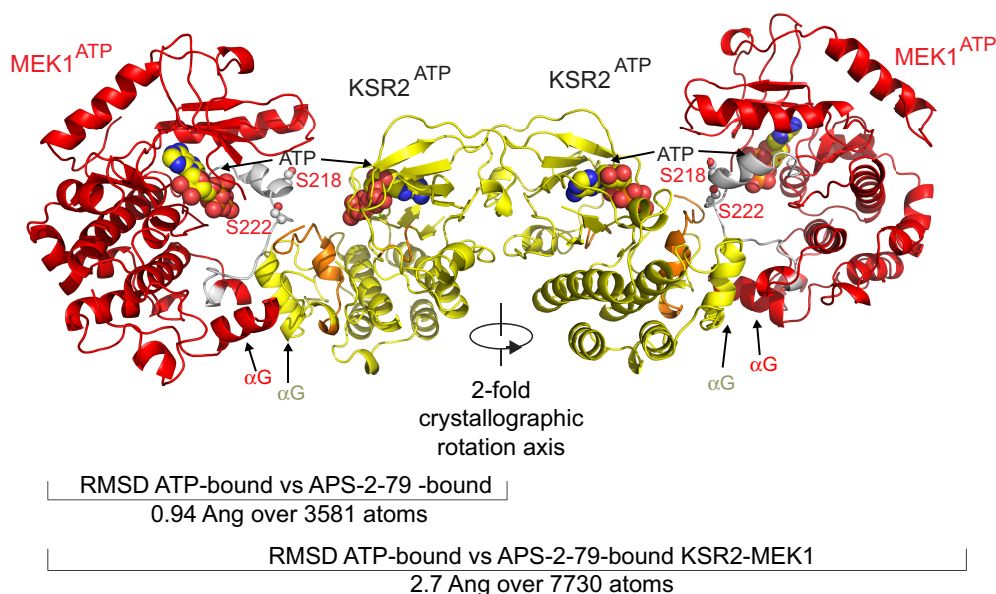
Extended Data Figure 4 | APS-2-79 activity is not dependent on KSR phosphorylation sites in MEK or direct RAF inhibition. **a**, APS-2-79 does not affect BRAF(V600E)-induced MAPK activation in cells. BRAF(V600E)-Flag was expressed for 24 h in 293H cells. Cells were then treated for 2 h with DMSO or 5 μ M of either APS-2-79, APS-3-77, or dabrafenib before collection and western blot analysis of phosphorylated MEK (MEK1/2(pSer218/pSer222)) and ERK (ERK1/2(pT202/pY204)). **b**, Removal of putative KSR phosphorylation sites in MEK (MEK(AAAA); S18A, T23A, S24A, S72A; ref. 7) neither hinders KSR-dependent MAPK signalling, nor the activity of APS-2-79. Co-expression of full-length KSR-Flag and wild-type MEK1-GFP or MEK(AAAA)-GFP leads to enhanced MAPK signalling within 293H cells as visualized by immunoblotting for phosphorylated MEK (MEK1/2(pSer218/pSer222)) and ERK (ERK1/2(pT202/pY204)). APS-2-79 impedes KSR-stimulated MAPK signalling within cells through wild-type and MEK(AAAA)

equally. Bars and error bars indicate pMEK and pERK intensity and standard deviations, respectively. Signals were normalized relative to lane 5. Error bars indicate the mean \pm s.d. ($n = 3$ biological replicates). *** $P < 0.0005$ by two-tailed unpaired t -testing. **c**, The dimer-deficient KSR(R718H) mutant, relative to wild-type KSR, is compromised in MEK-inhibitor-induced feedback. 293H cells were co-transfected with MEK-GFP and KSR-Flag or KSR(R718H)-Flag for 24 h and then treated with increasing concentrations of trametinib (range of 0.13 to 100 nM; threefold dilutions) for an additional 48 h. Cells were collected and analysed by western blot. **d**, Phospho-AMPK remains unchanged in HCT116 cells upon co-treatment with APS-2-79 and trametinib. HCT116 cells were treated with APS-2-79 and/or trametinib for 48 h. Phospho-AMPK (top), phospho-ERK(pERK), and total MEK (bottom) western blots are shown.

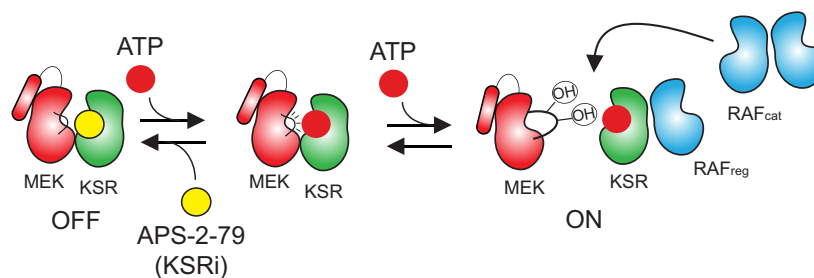
A. APS-2-79 Bound KSR2:MEK1 Complex



B. ATP Bound KSR2:MEK1 Complex



C.



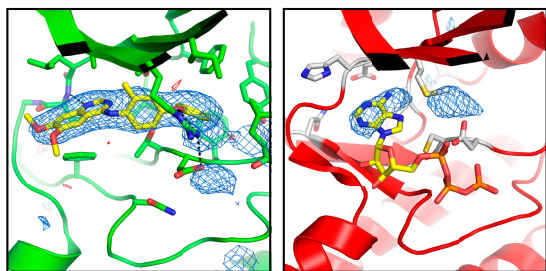
Extended Data Figure 5 | Higher order assembly of the KSR2-MEK1 complex bound to ATP or APS-2-79. **a**, Assembly of the KSR2-MEK1 heterodimer bound to APS-2-79. A crystal-packing two-fold symmetry axis of the asymmetric unit containing a single KSR2-MEK1 complex produces the heterotetramer. KSR2 bound to APS-2-79 is coloured green, and MEK1 is coloured red. The activation segments of KSR2 and MEK1 are coloured orange and white, respectively. The 'induced lock' (residues 809 to 814) within KSR2 is highlighted as orange, red and blue spheres. **b**, Assembly of the KSR2-MEK1 heterodimer bound to ATP as reported ref. 7 (PDB code: 2Y4I). A crystallographic two-fold rotation axis produces

the heterotetramer. r.m.s. deviation between the heterodimer and heterotetramers, respectively, of the ATP- and APS-2-79-bound KSR2-MEK1 complexes are listed below. **c**, A model for APS-2-79 function as a KSR2-targeted antagonist of MAPK signalling. APS-2-79 shifts the equilibrium of KSR2-MEK1 complexes so to populate the OFF state (left), and thereby antagonizes RAF dimerization and subsequent phosphorylation of KSR-bound MEK (far right). The model for RAF dimerization and MEK phosphorylation are adapted from ref. 7. In this model, the role of RAF_{cat} may be fulfilled by multiple active RAF-family kinases, such as C-RAF, bound within homo- or heterodimers of RAF-RAF or KSR-RAF, respectively.

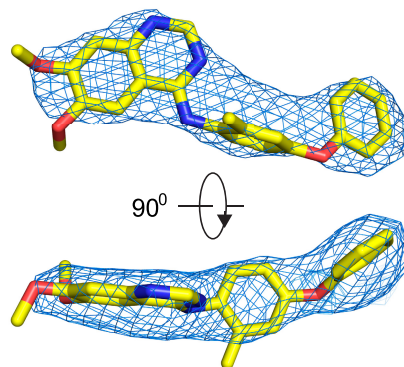
A.

KSR2 Active Site

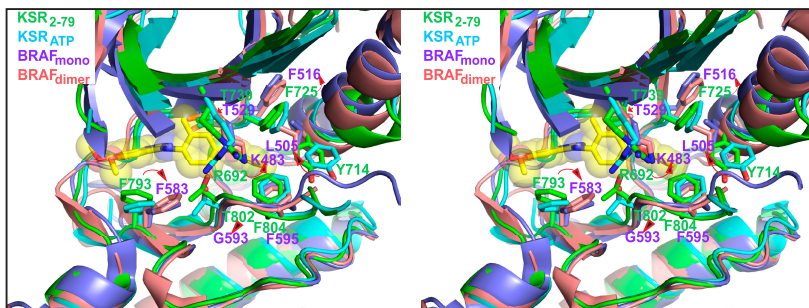
MEK1 Active Site



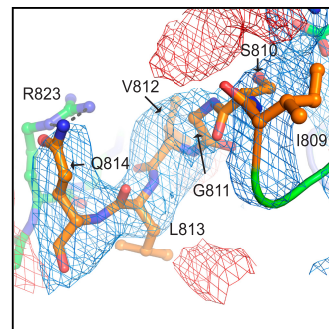
B.



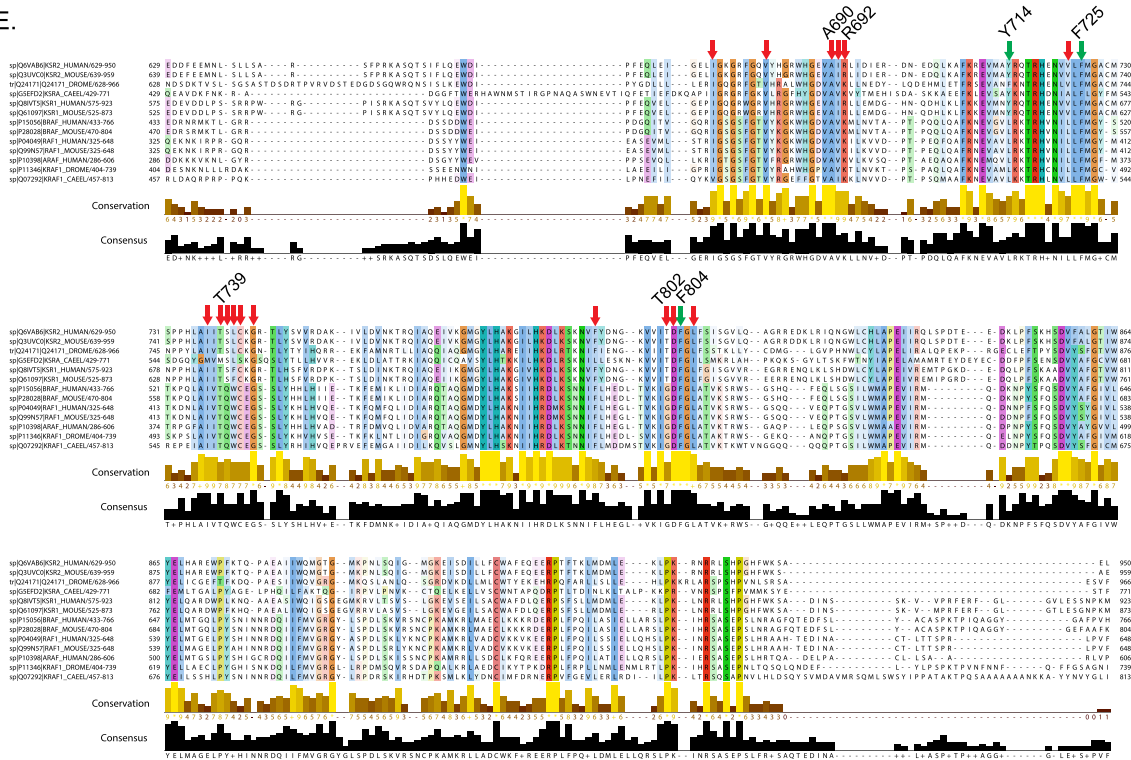
C.



D.



E.



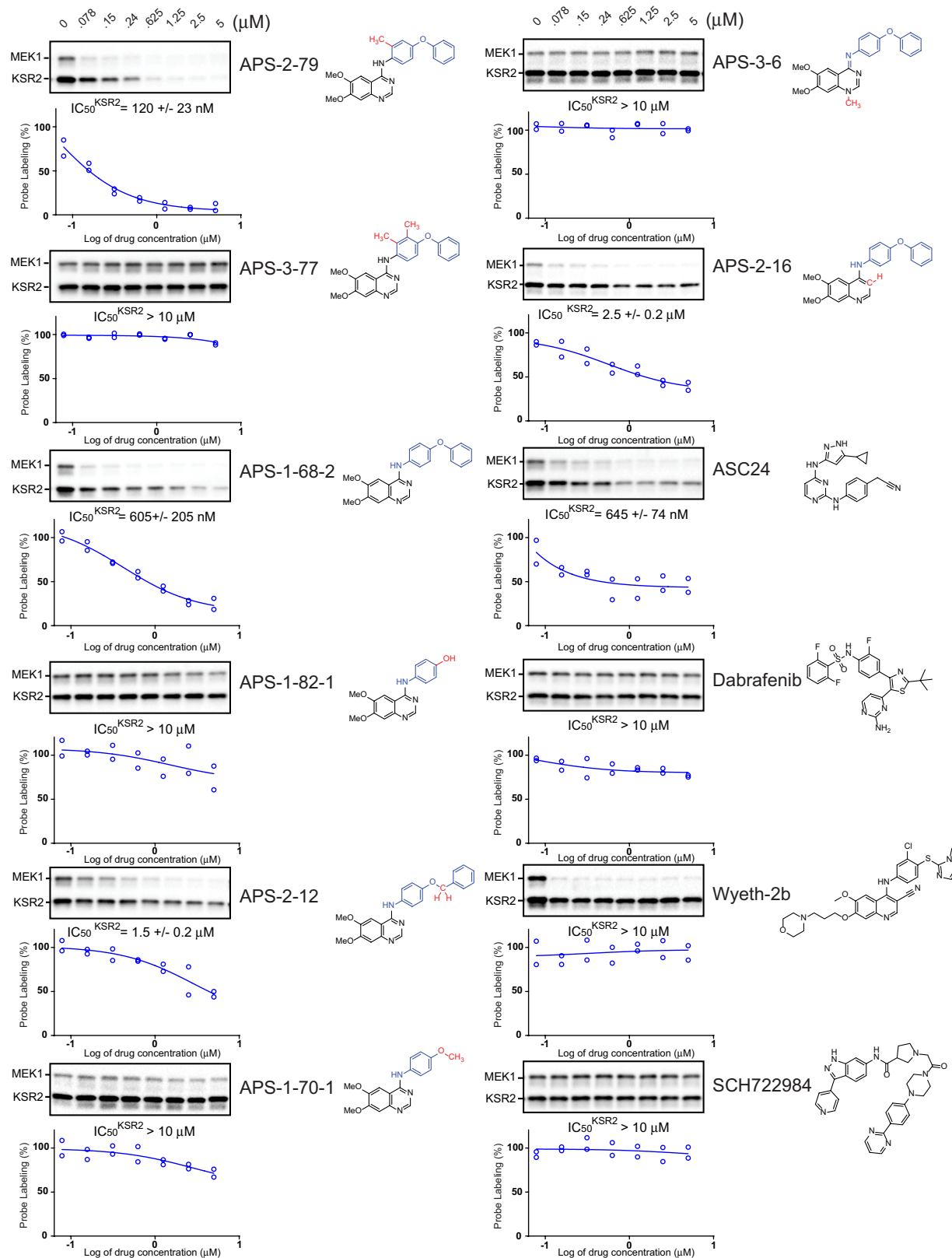
↓ APS-2-79 Contact Residues

↓ APS-2-79 Contact Residues That Form Pi-Stacking Interactions

Extended Data Figure 6 | See next page for caption.

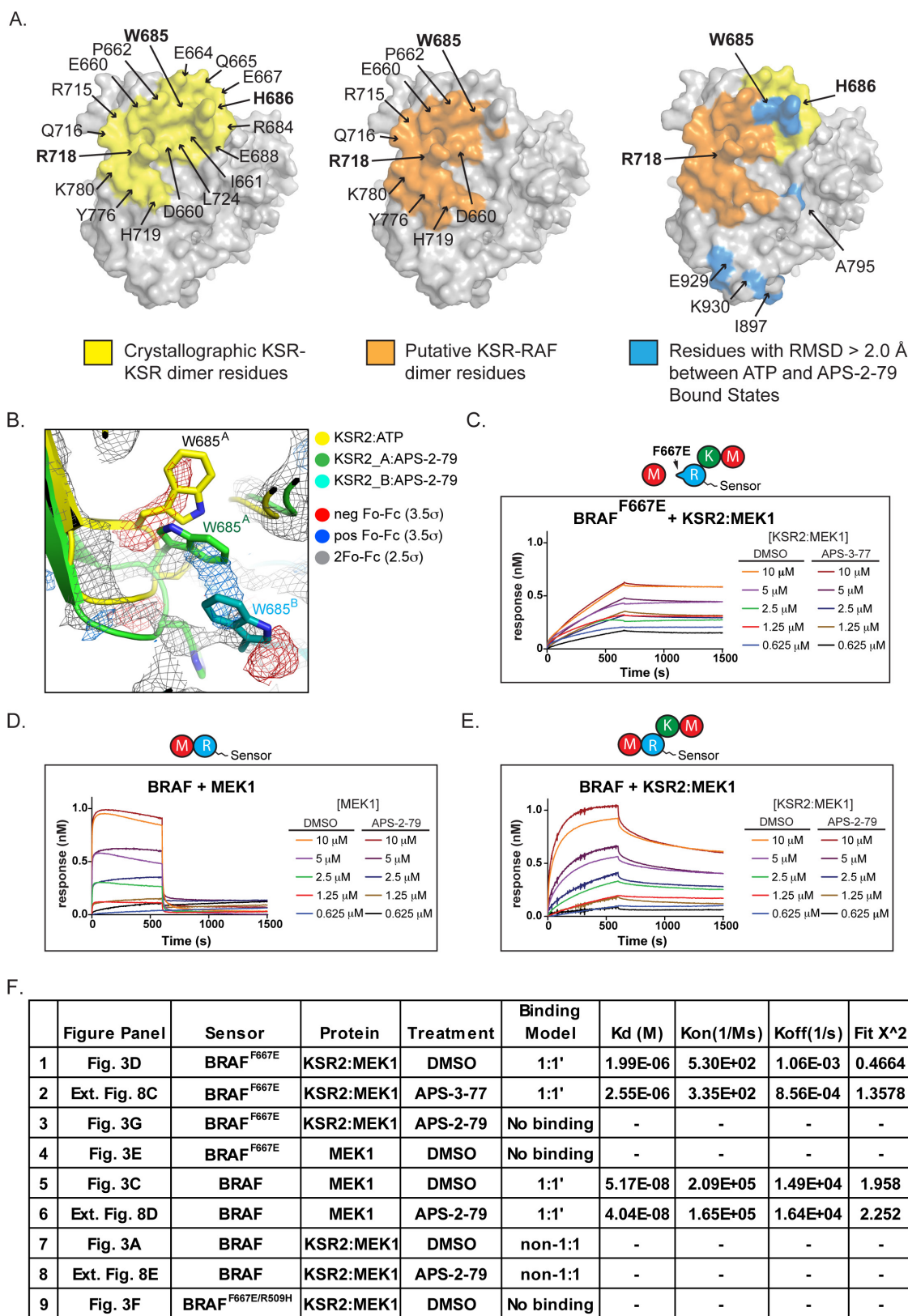
Extended Data Figure 6 | The APS-2-79 binding site within KSR2 and possible basis for KSR over RAF selectivity. **a**, APS-2-79 and ATP are overlaid in the KSR2 and MEK1 active sites, respectively. ATP was shown here to emphasize the MEK active site, but ATP was not included in the final model. Positive (blue) and negative (red) $F_o - F_c$ electron density maps, calculated before modelling of APS-2-79, are contoured at 3.5σ . Strong-positive-difference density within KSR2 supported modelling of APS-2-79 bound to KSR2 within the KSR2–MEK1 complex. **b**, Electron density map (blue mesh) for APS-2-79 (sticks) contoured at 4.5σ . Map represents positive difference density within the KSR2 active site before modelling of APS-2-79. **c**, Superposition of KSR2 (ATP- and APS-2-79-bound) with BRAF monomer (PDB code: 4W05) and BRAF dimer (PDB code: 3C4C) co-crystal structures reveals the possible bases for selectivity of APS-2-79 for KSR over RAF proteins. Residues within the APS-2-79 binding pocket that diverge between KSR and RAF proteins, but which are highly conserved within both sub-families are indicated with

arrows. Thr802 in KSR2, which is universally a Gly residue in all active RAF homologues, and also Phe516 and Phe793 in KSR2, which adopt distinct orientations from the equivalent Phe residues in RAF kinases, directly contact the biphenyl ether motif in APS-2-79. The T802G substitution, as well as the positional differences of the above-mentioned aromatic residues, would be predicted to reduce binding of active RAFs with APS-2-79. Another interaction that is probably favoured in KSR includes the contact mediated by the epsilon nitrogen of Arg692 with the –O– linker of the biphenyl motif; the placement of Arg692 is stabilized by Asp803 of the DFG motif. In RAF, the Arg-to-Lys substitution (Lys483 in subdomain II of BRAF), lacks the equivalent nitrogens to bond with both the –O– linker in APS-2-79 and the aspartate of the DFG motif. **d**, Positive (blue) and negative (red) $F_o - F_c$ electron density map contoured at $\pm 2.5\sigma$, before modelling of residues I809 to Q814 in KSR2, is shown. **e**, Sequence alignment of KSR and RAF proteins. Arrows highlight APS-2-79 contact residues.



Extended Data Figure 7 | *In vitro* ATP^{biotin} competition assays. Representative western blot images of *in vitro* ATP^{biotin} competition assays using recombinant KSR2–MEK1 and analogues reported in this study. Chemical structures are shown adjacent to assay blots. IC_{50} values

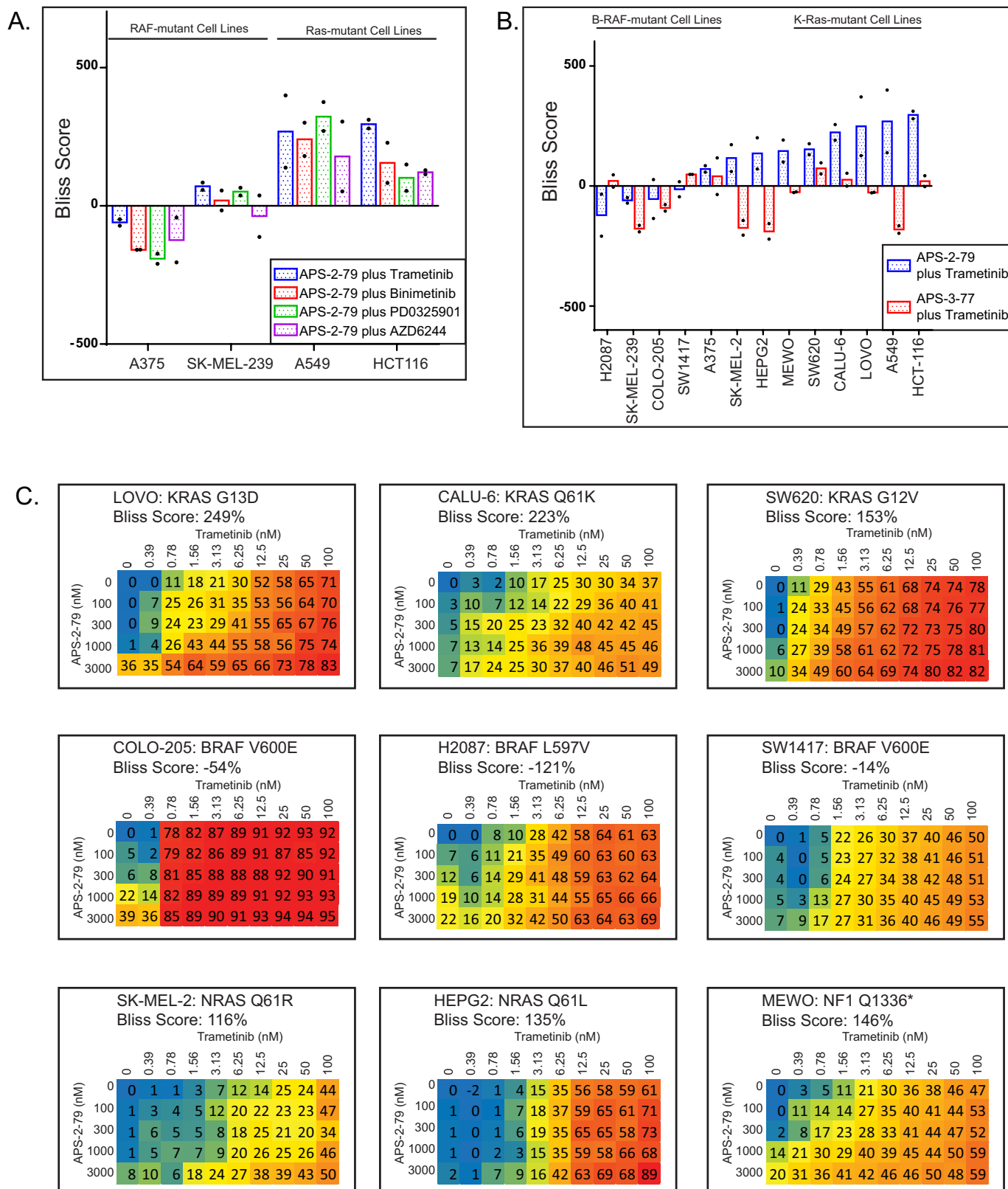
(mean \pm s.d.; $n = 2$ biological replicates) against ATP^{biotin} probe-labelling of KSR2 are listed below blots. Line graphs include data points from two biological replicates.



Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | Bio-layer interferometry binding data between BRAF and free MEK1 or the KSR2–MEK1 complex. **a**, Mapping of residues with a r.m.s. deviation of greater than 2.0 Å between the ATP- and APS-2-79-bound states of KSR2–MEK1 (right, blue), highlights alterations at contact residues Trp685 and His686 within the KSR–KSR homodimer (left, yellow) and KSR–RAF heterodimer (middle, orange) interfaces. **b**, Movement of Trp685–His686 within KSR2 between the ATP- and APS-2-79-bound states. A single protomer of KSR2 in the ATP-bound state (yellow), and both protomers (green and cyan) of the KSR2 dimer within the APS-2-79-bound state, are shown. Negative density around W685 and His686 in early-stage maps supported the conformational change in this loop between the ATP- and APS-2-79-bound states. **c**, The negative control compound APS-3-77 (25 µM) does not impact assembly of BRAF(F667E) and KSR2–MEK1. These assays were performed identically to the experiments in Fig. 3b–g. Coloured

curves indicate dose ranges of KSR2–MEK1 or MEK1 from 625 nM to 10 µM in the presence or absence of the indicated compounds. In all plots, association occurred from 0 to 660 s, and dissociation was monitored thereafter up to 1500 s. **d–e**, Biolayer interferometry of wild-type BRAF with MEK1 and KSR2–MEK1 in the presence of DMSO and 25 µM APS-2-79. These assays were performed identically to the experiments in Fig. 3b–g. Coloured curves indicate dose ranges of KSR2–MEK1 or MEK1 from 625 nM to 10 µM in the presence or absence of the indicated compounds. In all plots, association occurred from 0 to 660 s, and dissociation was monitored thereafter up to 1500 s. **f**, Table summary of BLI data in this figure and Fig. 3b–h. K_d , K_{on} , and K_{off} values represent the mean and s.e.m. measurements derived from global fitting of 5 binding curves. χ^2 and R^2 describe experimental and model data correlations; <3 and above 0.95, respectively, indicate good fits.

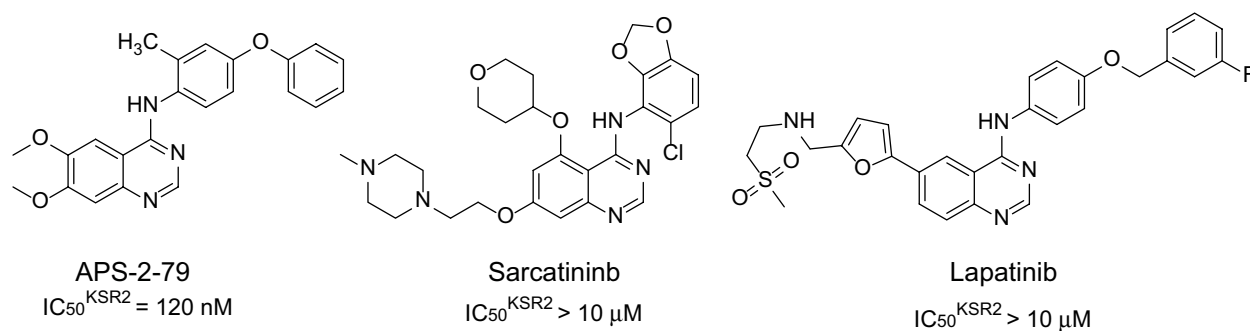


Extended Data Figure 9 | See next page for caption.

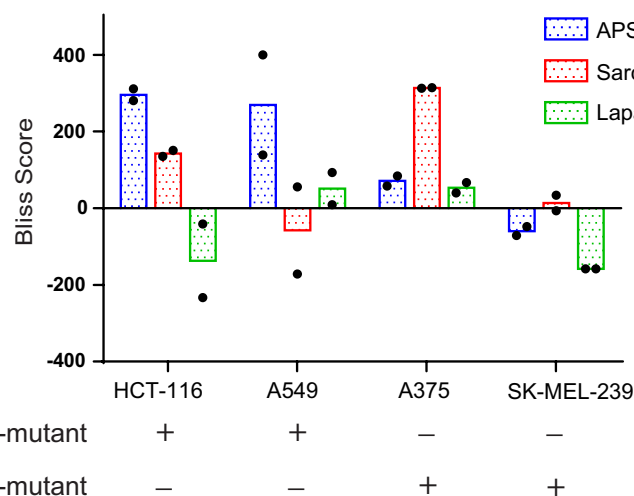
Extended Data Figure 9 | KSRi binder APS-2-79 synergizes with trametinib in Ras-mutant cells. **a**, Average Bliss score of the combination of trametinib, binimetinib, PD0325901, or AZD6244 with APS-2-79 in the Ras-mutant cell lines HCT116 and A549 versus the RAF-mutant cell lines A375 and SK-MEL-239. Full combination matrices of APS-2-79 (range: 100 nM to 3 μ M in threefold dilutions) with trametinib (range: 0.01–100 nM in threefold dilutions), binimetinib (range: 0.1–10 μ M in threefold dilutions), PD0325901 (range: 0.1–10 μ M in threefold dilutions), and AZD6244 (range: 0.1–10 μ M in threefold dilutions). Bars represent the mean Bliss scores calculated from two biological replicates of the depicted concentration matrices; points represent each calculated score. **b**, Average Bliss scores of APS-2-79 or APS-3-77 in combination with trametinib in RAF-mutant, RAS-mutant cell lines. SK-MEL-2 and HepG2

are N-Ras-mutant cell lines, and MEWO is a NF1-mutant cell line. Bars represent the mean Bliss scores calculated from two biological replicates of the depicted concentration matrices; points represent each calculated score. **c**, Complete cell viability analysis of APS-2-79 (range: 100–3,000 nM in threefold dilutions) plus trametinib (range: 0.01–100 nM in threefold dilution) over a full concentration matrix in the Ras-mutant LOVO, CALU-6, SW620, SK-MEL-2, and HEPG2 cell lines, the RAF-mutant COLO-205, H2087, and SW1417 cells, and the NF1-mutant MEWO cell line. Numbers listed within synergy matrices represent percentage of growth inhibition relative to DMSO control and are the mean of two biological replicates. Bliss scores represent the mean calculated from two biological replicates of the depicted concentration matrices.

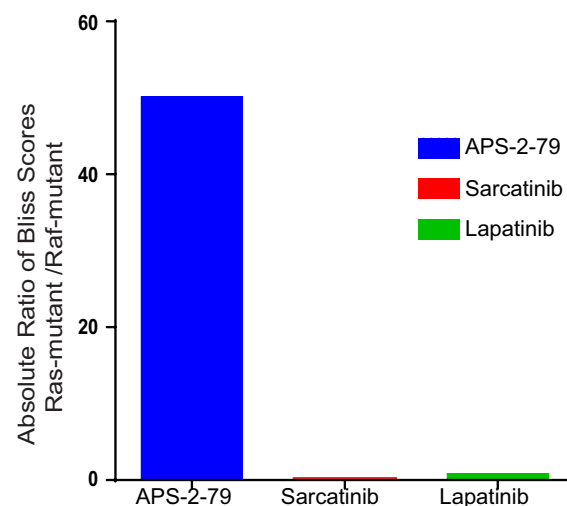
A.



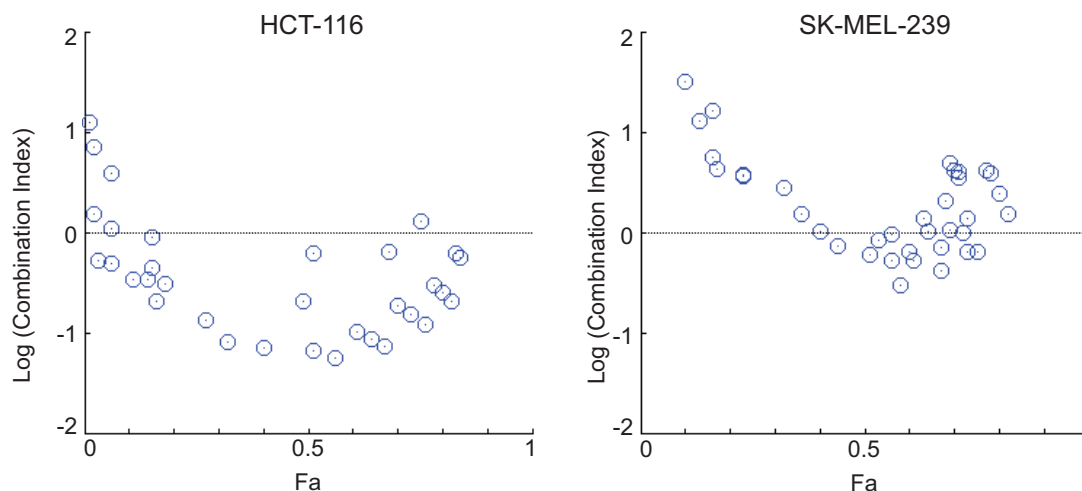
B.



C.



D.



Extended Data Figure 10 | APS-2-79 synergizes with trametinib specifically in Ras-mutant cells compared to the HER-family and SRC-family inhibitors lapatinib and sarcatinib. **a**, Chemical structures of APS-2-79 and quinazoline-containing kinase inhibitors sarcatinib and lapatinib. The primary targets for sarcatinib and lapatinib are c-Src and Her2, respectively³⁰. IC₅₀ values against ATP^{biotin} probe-labelling of KSR2 are listed below structures. **b**, Bliss score analysis of HCT-116, A549, A375, and SK-MEL-239 cells treated with APS-2-79, sarcatinib, or lapatinib (range: 100–3,000 in threefold dilutions) in combination with trametinib (range: 0.01–100 in threefold dilution). Bars represent the mean Bliss

scores calculated from two biological replicates; points represent each calculated score. **c**, Absolute Bliss score of the indicated drugs in combination with trametinib in Ras-mutant relative to RAF-mutant cell lines demonstrates selective synergy in Ras-mutant cell lines for APS-2-79 compared to sarcatinib and lapatinib. **d**, log of the combination index graphs of APS-2-79 in combination with trametinib in HCT-116 versus SK-MEL-239 cells as compared to the fractional effect. Negative combination index over a broad fractional effect range within HCT-116, but not SK-MEL-239, indicates strong synergy.

28. Mallon, R. *et al.* Identification of 4-anilino-3-quinolinecarbonitrile inhibitors of mitogen-activated protein/extracellular signal-regulated kinase 1 kinase. *Mol. Cancer Ther.* **3**, 755–762 (2004).
29. Morris, E. J. *et al.* Discovery of a novel ERK inhibitor with activity in models of acquired resistance to BRAF and MEK inhibitors. *Cancer Discovery* **3**, 742–750 (2013).
30. Anastassiadis, T., Deacon, S. W., Devarajan, K., Ma, H. & Peterson, J. R. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, 1039–1045 (2011).

Structural basis for inhibition of a voltage-gated Ca^{2+} channel by Ca^{2+} antagonist drugs

Lin Tang^{1,2}, Tamer M. Gamal El-Din¹, Teresa M. Swanson¹, David C. Pryde³, Todd Scheuer¹, Ning Zheng^{1,2§} & William A. Catterall^{1§}

Ca^{2+} antagonist drugs are widely used in therapy of cardiovascular disorders^{1,2}. Three chemical classes of drugs bind to three separate, but allosterically interacting, receptor sites on $\text{Ca}_v1.2$ channels, the most prominent voltage-gated Ca^{2+} (Ca_v) channel type in myocytes in cardiac and vascular smooth muscle^{3–9}. The 1,4-dihydropyridines are used primarily for treatment of hypertension and angina pectoris and are thought to act as allosteric modulators of voltage-dependent Ca^{2+} channel activation, whereas phenylalkylamines and benzothiazepines are used primarily for treatment of cardiac arrhythmias and are thought to physically block the pore^{1,2}. The structural basis for the different binding, action, and therapeutic uses of these drugs remains unknown. Here we present crystallographic and functional analyses of drug binding to the bacterial homotetrameric model Ca_v channel Ca_vAb , which is inhibited by dihydropyridines and phenylalkylamines with nanomolar affinity in a state-dependent manner. The binding site for amlodipine and other dihydropyridines is located on the external, lipid-facing surface of the pore module, positioned at the interface of two subunits. Dihydropyridine binding allosterically induces an asymmetric conformation of the selectivity filter, in which partially dehydrated Ca^{2+} interacts directly with one subunit and blocks the pore. In contrast, the phenylalkylamine Br-verapamil binds in the central cavity of the pore on the intracellular side of the selectivity filter, physically blocking the ion-conducting pathway. Structure-based mutations of key amino-acid residues confirm drug binding at both sites. Our results define the structural basis for binding of dihydropyridines and phenylalkylamines at their distinct receptor sites on Ca_v channels and offer key insights into their fundamental mechanisms of action and differential therapeutic uses in cardiovascular diseases.

Ca_v1 channels are composed of a complex of a pore-forming $\alpha1$ subunit associated with β , γ , and $\alpha2\delta$ subunits^{1,10}. The $\alpha1$ subunits contain four homologous domains with six transmembrane segments in each^{11,12}. Transmembrane segments S1–S4 form the voltage-sensing module, and S5, S6 and the intervening P-loop form the pore¹. The overall architecture of the mammalian skeletal muscle $\text{Ca}_v1.1$ channel was recently elucidated at a resolution of $\sim 4\text{--}6\text{ \AA}$ by cryo-electron microscopy¹³. However, higher-resolution structural analysis of mammalian Ca_v channels has not yet been achieved. The bacterial voltage-gated Na^+ channel NaChBac and its relatives are homotetrameric proteins composed of four identical subunits, each analogous to one domain of a mammalian voltage-gated Na^+ or Ca^{2+} channel^{14,15}. These bacterial channels probably represent the evolutionary ancestors of both mammalian channel families. The structures of bacterial Na^+ channels have been determined at high resolution by X-ray crystallography in pre-open¹⁶ and inactivated^{17,18} states. Moreover, the structural basis for Ca^{2+} conductance and selectivity has been elucidated at atomic resolution through studies of Ca_vAb , a site-directed mutant of Na_vAb with full Ca^{2+} channel function¹⁹. We have used derivatives of Ca_vAb

(see Methods) to define receptor sites and mechanisms of action of Ca^{2+} antagonist drugs at atomic resolution.

Ca_vAb was inhibited by amlodipine with high affinity (Fig. 1a–c). No inhibition was observed during single depolarizations, indicating that amlodipine does not enter the open pore and block it (Fig. 1a). However, inhibition increased progressively during trains of depolarizations, reflecting increased binding affinity for the activated and/or inactivated states of Ca_vAb (Fig. 1b). After a train of 20 depolarizing pulses, the half-maximum inhibitory concentration (IC_{50}) for inhibition by amlodipine was 10 nM (Fig. 1c). This affinity was surprisingly high, considering the evolutionary distance between Ca_vAb and mammalian $\text{Ca}_v1.2$ channels, which have IC_{50} values from 0.3 nM to 1 μM for various dihydropyridines²⁰.

Photoaffinity labelling and site-directed mutagenesis suggest that dihydropyridines bind to a receptor site at the interface of homologous domains III and IV and the adjacent pore module in domain III in $\text{Ca}_v1.2$ channels^{4–7,21,22}. In Ca_vAb , four identical subunits form a homotetramer (Fig. 1d)¹⁹. The structure of the amlodipine– Ca_vAb complex reveals the antagonist bound on the outer, lipid-facing surface of the pore module in the intersubunit crevice formed by neighbouring tilted S6 helices and the P-helix of the selectivity filter (Fig. 1d, e, yellow sticks). Despite the homotetrameric structure of Ca_vAb , only a single drug-binding site per tetramer is occupied, suggesting that drug-induced conformational changes prevent occupancy of more than one site. Amino-acid residues Y195, I199, F171, Y168 and F167 form a hydrophobic pocket for interaction with amlodipine (Fig. 1f). The dihydropyridine ring is sandwiched between Y195 of S6 and F167 of the P-loop. F171 and I199 of S6 form the bottom of the cleft that accommodates the bound drug. Mutations of I199 (for example, I199S) had minimal effects on Ca_vAb function (Extended Data Fig. 1), but markedly reduced the affinity for amlodipine ($\text{IC}_{50} = 112\text{ nM}$; Fig. 1c).

Nimodipine inhibited Ca_vAb like amlodipine, but its IC_{50} was 100 nM (Fig. 2a–c). Nimodipine binds to the same site as amlodipine (Fig. 2d, e and Extended Data Fig. 2a, b). The substitution I199S increased the IC_{50} for nimodipine from 100 nM to 5.7 μM (Fig. 2c), and W195Y increased it to 508 nM (Extended Data Fig. 3). The experimental Br-dihydropyridine derivative UK-59811 inhibited Ca_vAb with $\text{IC}_{50} = 194\text{ nM}$ (Extended Data Fig. 4) and bound in a similar position (Fig. 2f and Extended Data Fig. 2a, c). Anomalous scattering density from its Br atom further confirmed the location of the dihydropyridine-binding site at the interface between the S6 segments of two adjacent subunits surrounded by Y195, F171, F167, and I199 (Fig. 2f, green mesh). High-resolution structures of Ca_vAb revealed 16 molecules of bound lipid per tetramer¹⁹. Without drugs, we found a single molecule of DMPC lipid aligned in the dihydropyridine-binding pocket with its polar headgroup facing the extracellular side and its long hydrocarbon tails projecting deep into the crevice formed by neighbouring S6 helices (Fig. 2g and Extended Data Fig. 2d). Thus, our

¹Department of Pharmacology, University of Washington, Seattle, Washington 98195-7280, USA. ²Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195-7280, USA. ³Curadev Pharma, Discovery Park, Sandwich, Kent CT14 9FF, UK. §These authors jointly supervised this work.

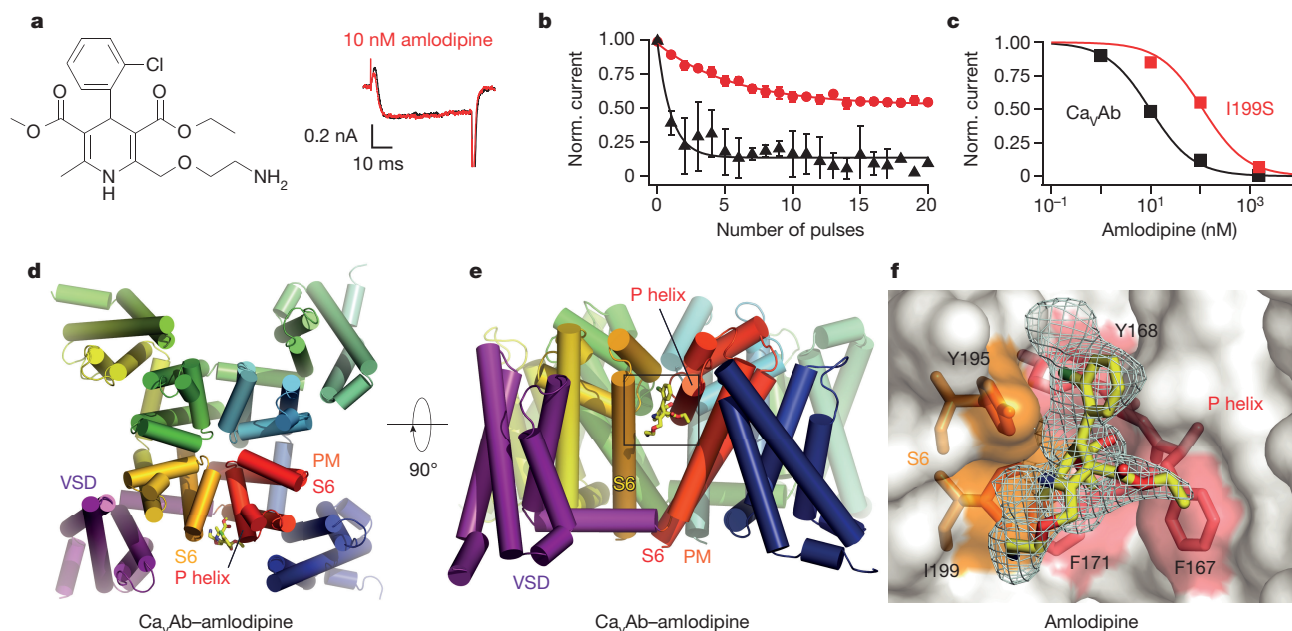


Figure 1 | Structural basis for inhibition of $\text{Ca}_\text{V}\text{Ab}$ by amlodipine.

a, Amlodipine structure. Ba^{2+} currents for 0 nM (black) and 10 nM (red) amlodipine during depolarization from -120 mV to 0 mV. **b**, State-dependent block by amlodipine after 50-ms pulses at 1 Hz from -120 mV to 0 mV (10 nM, circles; 100 nM, triangles; mean \pm s.e.m.; $n = 3-5$). **c**, Inhibition by amlodipine. Data were fit by the Hill equation with $n_{\text{H}} = 1$. $\text{Ca}_\text{V}\text{Ab}$: $\text{IC}_{50} = 10 \pm 0.4$ nM; $\text{Ca}_\text{V}\text{Ab}$ I199S: $\text{IC}_{50} = 112 \pm 10$ nM; $n = 3-5$;

mean \pm s.e.m. **d**, Structure of $\text{Ca}_\text{V}\text{Ab}$ (top view in cylinders)

binding amlodipine (yellow sticks). PM, pore module; VSD, voltage-sensing domain. **e**, $\text{Ca}_\text{V}\text{Ab}$ with bound amlodipine in side view. **f**, Dihydropyridine-binding pocket of $\text{Ca}_\text{V}\text{Ab}$ with the F_0-F_c electron density map (2.5σ , cyan) and amlodipine (yellow sticks). $\text{Ca}_\text{V}\text{Ab}$ residues contacted by amlodipine are highlighted in colours and labelled.

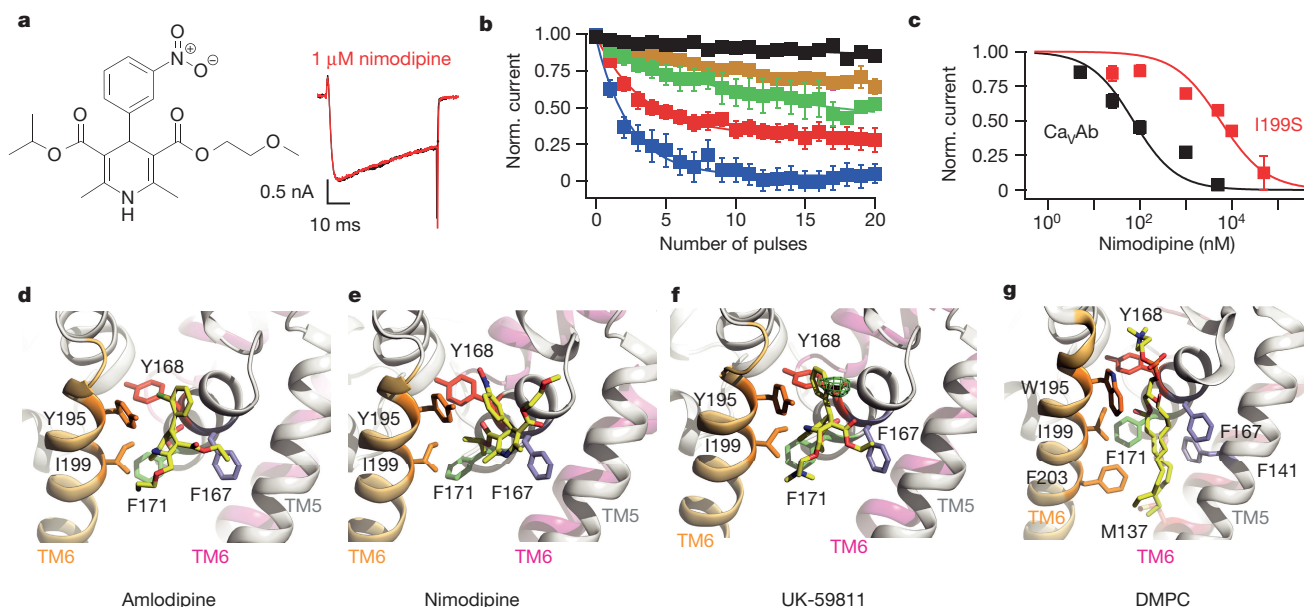


Figure 2 | Inhibition of $\text{Ca}_\text{V}\text{Ab}$ by dihydropyridine binding at a lipid site.

a, Nimodipine structure. Current records as in Fig. 1a. **b**, State-dependent block by nimodipine as in Fig. 1b: 5 nM (black), 25 nM (brown), 100 nM (green), 1 μM (red), and 5 μM (blue); mean \pm s.e.m.; $n = 3-14$. **c**, Inhibition by nimodipine as in Fig. 1c. $\text{Ca}_\text{V}\text{Ab}$: $\text{IC}_{50} = 100 \pm 9$ nM; $\text{Ca}_\text{V}\text{Ab}$

I199S: $\text{IC}_{50} = 5.7 \pm 0.6$ μM ; $n = 3-14$; mean \pm s.e.m. **d**, Amlodipine (yellow sticks) bound to $\text{Ca}_\text{V}\text{Ab}$. S5 and S6 helices in ribbons; residues surrounding amlodipine in sticks. **e**, Nimodipine bound to $\text{Ca}_\text{V}\text{Ab}$. **f**, UK-59811 bound to $\text{Ca}_\text{V}\text{Ab}$. Anomalous scattering density (3σ , green mesh) for Br in UK-59811. **g**, DMPC lipid in the drug-free dihydropyridine-binding site in yellow sticks.

structures reveal that dihydropyridine binding displaces an endogenous lipid molecule from their common binding site on $\text{Ca}_\text{V}\text{Ab}$.

In the absence of dihydropyridines, the $\text{Ca}_\text{V}\text{Ab}$ structure has fourfold symmetry around the pore axis¹⁹. Four lipid molecules are found in the central cavity, occupying fenestrations that connect to the exterior of the channel (Fig. 3a). Binding of dihydropyridines to $\text{Ca}_\text{V}\text{Ab}$ rearranges the quaternary structure and breaks the fourfold symmetry (Fig. 3c, e, g, compare shaded cross-sections; see Supplementary Discussion of

asymmetry induced by drug binding). With drug bound, the four lipid molecules in the central cavity lose their symmetric spatial organization, and the fenestration closest to the drug-binding site is no longer occupied by a lipid chain.

By introducing asymmetry, dihydropyridine binding triggers allosteric changes at the selectivity filter of $\text{Ca}_\text{V}\text{Ab}$ and alters binding of the substrate ion. There are three Ca^{2+} -binding sites in the $\text{Ca}_\text{V}\text{Ab}$ selectivity filter: two high-affinity sites (Sites 1 and 2) followed by one

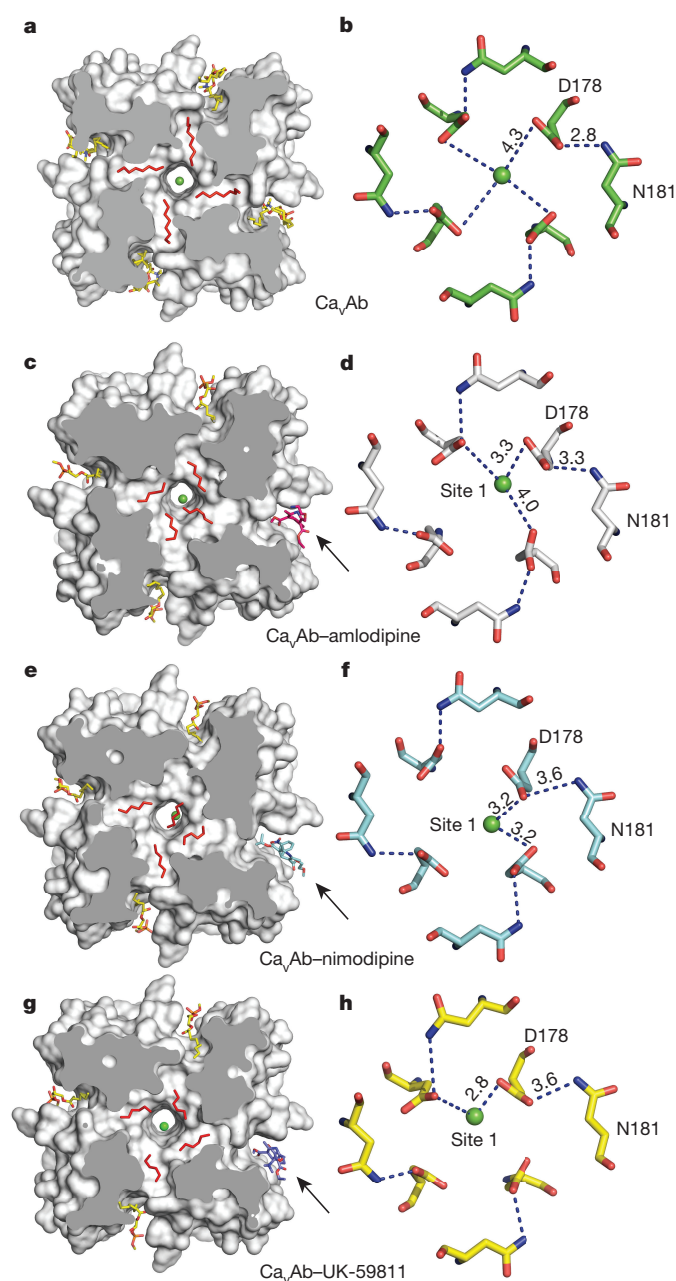


Figure 3 | Dihydropyridine binding allosterically modifies Ca^{2+} binding in the selectivity filter. **a**, Outward view. Four symmetrical lipids (red sticks) occupy fenestrations in Ca_vAb without dihydropyridine. Four additional lipids bind to the side of the pore module (yellow sticks). **b**, Top view. Site 1 with hydrated Ca^{2+} (green) coordinated directly by D178 and indirectly by N181 on extracellular end of the selectivity filter. **c**, Amlodipine binding (magenta sticks) induces asymmetry and causes rearrangement of lipids (red sticks). **d**, Top view. Site 1 with partially dehydrated Ca^{2+} and direct interaction with D178 due to binding of amlodipine. **e**, Binding of nimodipine (cyan sticks) induces asymmetry and reorganizes bound lipid. **f**, Partially dehydrated Ca^{2+} binds at site 1 with coordination distance of 3.2 Å to carboxylate side chains of D178. **g**, Binding of UK-59811 (blue sticks) to the dihydropyridine binding site induces asymmetry and reorganizes bound lipid. **h**, Ca^{2+} binds at Site 1 with coordination distance of 2.8 Å to a carboxylate side chain of D178.

lower-affinity site (Site 3) arranged sequentially from its extracellular to intracellular end¹⁹. Without drug, Ca^{2+} binds near the central axis of the pore in a fully hydrated state, coordinated symmetrically by four D178 carboxylate side chains (Fig. 3b)¹⁹. With dihydropyridines bound, Ca^{2+} binds to Site 1 asymmetrically in a partially dehydrated

state—significantly off the central axis of the pore and closer to one or two D178 carboxylate groups at a distance of 2.8–3.3 Å (Fig. 3d, f, h and Extended Data Fig. 5a–d). This binding distance suggests direct interaction of bound Ca^{2+} with the carboxylate side chain (Supplementary Discussion). In contrast, binding of Ca^{2+} at Site 2 is unchanged (data not shown). The anomalous scattering density of Ca^{2+} confirms its off-axis location in Site 1 and on-axis location in Site 2 (Extended Data Fig. 5e, f).

Studies with quaternary phenylalkylamine analogues revealed that these drugs inhibit $\text{Ca}_v1.2$ channels only after cytoplasmic application, and that drug binding is increased by repetitive depolarization to open the pore^{2,23}. It was therefore concluded that tertiary phenylalkylamines such as verapamil penetrate the membrane in uncharged form, are re-protonated in the cytosol, and block the $\text{Ca}_v1.2$ channel by entering the intracellular mouth of the open pore in their protonated form and binding to their receptor site^{2,23}. Photoaffinity labelling and site-directed mutagenesis revealed that the phenylalkylamine receptor site is formed by S6 segments in domains III and IV of $\text{Ca}_v1.2$ channels, consistent with drug binding in the pore^{4–7,24,25}.

When Br-verapamil was perfused at –120 mV, the first depolarization to 0 mV showed progressive reduction of the current during the pulse (Fig. 4a). This profile supports a pore-blocking mechanism, in which the drug progressively enters and blocks the open pore. Repetitive depolarizing stimuli increased inhibition of Ca_vAb by Br-verapamil (Fig. 4b), yielding IC_{50} values of 810 nM for Br-verapamil (Fig. 4c, blue squares) and 475 nM for verapamil (Extended Data Fig. 6a, b) at steady state. The action of these drugs is strikingly state-dependent: the IC_{50} for Br-verapamil in the resting state is 24 μM , 30-fold higher than observed after a train of depolarizing stimuli (Fig. 4c, blue circles).

Our crystal structures revealed a single molecule of Br-verapamil bound in the central cavity on the intracellular side of the ion selectivity filter (Fig. 4d, e; see Supplementary Discussion of asymmetry induced by drug binding). The bound drug is oriented with its characteristic positively charged tertiary amino group facing in the extracellular direction pointing towards Site 3 in the selectivity filter. In this position, the bound phenylalkylamine would physically block the pore. The distance between the tertiary amino group and Ca^{2+} coordinated by the carbonyls of L176 is 5 Å. The methoxy groups in the aromatic rings are located close to the inner end of the fenestrations, surrounded by T206, M209 of the neighbouring subunit and T175, M174, L176 of the selectivity filter (Fig. 4f). The two aromatic rings of Br-verapamil interact with T206 residues from two neighbouring S6 helices (Fig. 4f). A view from the intracellular side shows that Br-verapamil binds closer to two subunits on one side of the pore (Fig. 4f). The anomalous scattering from Br-verapamil further defines the position of the aromatic ring that is farther from the amino group and confirms its interaction with T206 (Fig. 4e, green mesh). Mutations in T206 impair inactivation of Ca_vAb (Extended Data Fig. 6c, e) and markedly reduce the affinity for Br-verapamil. For example, the conservative mutation T206S increases the IC_{50} for state-dependent inhibition from 810 nM to 24 μM (Fig. 4c, red squares) and the IC_{50} for resting state inhibition of Ca_vAb from 24 μM to 115 μM (Fig. 4c, red circles). The effects of these mutations on both resting and state-dependent block confirm that there is a direct interaction between the drug and T206. These results define the receptor site for pore block by phenylalkylamines at high resolution. Similar to the dihydropyridine-binding site, the phenylalkylamine-binding site is also occupied by lipid molecules in the absence of the drug.

At concentrations above 1 μM , dihydropyridines inhibit voltage-gated Na^+ channels in a manner consistent with pore block²⁶. At the high drug concentrations used in our crystallization studies, we found binding of UK-59811 (Fig. 4g–i) and other dihydropyridines in the pore of Ca_vAb . The anomalous scattering density of its Br places the dihydropyridine ring deep in the central cavity where it forms hydrophobic contacts with two neighbouring subunits (Fig. 4h, green mesh). Compared to Br-verapamil, UK-59811 bound more towards

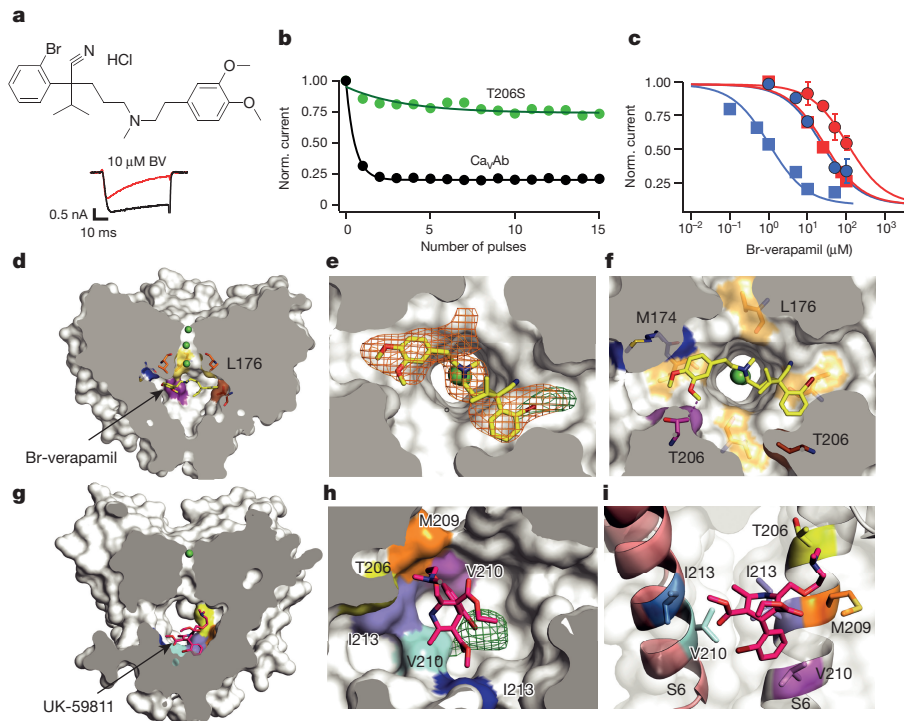


Figure 4 | State-dependent inhibition by pore block with Br-verapamil and UK-59811. **a**, Br-verapamil. Ba²⁺ current records for Ca_VAb with 0 μM (black) and 10 μM (red) during the depolarizing pulse. **b**, State-dependent block of Ca_VAb ($n = 7$) and Ca_VAb T206S ($n = 3$) at 10 μM during trains of depolarizations at 1 Hz from -120 mV to 0 mV. The error bars for all the data points on this graph are too small to be visible. **c**, Inhibition by Br-verapamil for Ca_VAb and Ca_VAb T206S at V = -120 mV and following trains of depolarizations as in **b**. Ca_VAb: resting state block, blue circles, IC₅₀ = 24 ± 1.6 μM; state-dependent block, blue squares, IC₅₀ = 810 ± 80 nM. Ca_VAb T206S: resting state block, red circles,

IC₅₀ = 115 ± 3.2 μM; state-dependent block, red squares, IC₅₀ = 24 ± 0.8 μM; $n = 3-11$; mean ± s.e.m. **d**, Side view of the pore module sectioned through the selectivity filter with Br-verapamil bound (yellow sticks). Ca²⁺, green spheres. **e**, F_o-F_c electron density (2.5σ, orange mesh) and anomalous scattering density (3σ, green mesh) for Br defines location of Br-verapamil. **f**, The two aromatic rings of verapamil are close to T206 of adjacent subunits. **g**, UK-59811 (red sticks) binds with its dihydropyridine ring deep in the central cavity. **h**, Anomalous scattering density (3.5σ, green mesh) of Br in UK-59811. **i**, S6 segments with residues surrounding UK-59811 in sticks.

the intracellular base of the central cavity and was located closer to one subunit (Fig. 4g-i). Low-affinity block of Na⁺ channels by dihydropyridines bound in this site may contribute to cardiac arrhythmias caused by toxic overdoses of these drugs.

Overall, our results provide a structural basis for understanding how dihydropyridines and phenylalkylamines bind at two distinct, but allosterically coupled receptor sites on Ca_V1.2 channels and have different efficacy for treatment of hypertension and angina pectoris versus cardiac arrhythmias⁵⁻⁷. Consistent with photoaffinity-labelling and site-directed mutagenesis⁵⁻⁷, dihydropyridines bind on the outer, lipid-facing surface of the pore module at the interface between two subunits of Ca_VAb, in analogy with their proposed site of action between domains III and IV of Ca_V1.2 channels^{4,21}. Their binding site is exposed to the extracellular side of the membrane, but not to the intracellular side. These structural results reveal why charged dihydropyridines are ineffective when applied intracellularly²⁷, and they are consistent with location of the drug-binding site ~11-14 Å from the outer surface of the lipid bilayer as inferred from studies of charged derivatives of amlodipine with hydrophobic linkers of increasing length²⁸. These comparisons reveal a close analogy between the site of dihydropyridine binding in our crystal structures of Ca_VAb and the expectations from studies of Ca_V1.2 channels, but the exact position of the drug-binding site in Ca_VAb is approximately one helical turn towards the extracellular side from the amino-acid residues implicated in dihydropyridine binding by studies of Ca_V1.2 channels (Extended Data Fig. 7). This difference may reflect the great evolutionary distance between Ca_VAb and mammalian Ca_V channels and/or indirect allosteric effects of mutations studied in Ca_V1 channels.

Binding of a single dihydropyridine to Ca_VAb induces a conformational change that alters the fourfold symmetry of the quaternary structure and induces changes in the three unoccupied dihydropyridine-binding sites that may prevent drug occupancy (Extended Data Fig. 8). Drug binding also disrupts the symmetry of the ion selectivity filter, allowing direct coordination of Ca²⁺ by carboxylate side chains. This conformational change is mediated in part by an altered pattern of hydrogen bonds formed by N181 in the subunit binding the dihydropyridine (Fig. 3). These structural results correlate closely with ligand-binding studies of Ca_V1.2 channels, which suggested that dihydropyridines induce high-affinity Ca²⁺ binding and block of the pore^{29,30}. Our structural studies reveal exactly how dihydropyridines act as indirect allosteric blockers of the pore of Ca²⁺ channels. Dihydropyridine binding to Ca_V1.2 channels is voltage-dependent because of the high affinity for the inactivated state^{1,5-7}. In a remarkable parallel, dihydropyridine binding causes a conformational change to an asymmetric pore structure in Ca_VAb, which is similar to the asymmetry induced in inactivated states of the parent Na_VAb channel¹⁷ and its relative Na_VRh¹⁸. Dihydropyridine binding may induce a similar asymmetric, Ca²⁺-blocked state of Ca_V1.2 channels and thereby enhance their inactivation, allowing selective inhibition in persistently depolarized cells. This mechanism underlies the use of dihydropyridines in treatment of hypertension and angina pectoris, in which vascular smooth muscle cells of resistance vessels are persistently depolarized, and their Ca_V1.2 channels are selectively inhibited by dihydropyridines.

The phenylalkylamine receptor site was localized to the S6 segments in domains III and IV of Ca_V1.2 channels by photoaffinity labelling and mutational analysis, and it was proposed that the amino-acid side chains involved in drug binding point towards the lumen of the

pore^{4–6,24,25}. Our structural results correlate precisely with this expectation and reveal the exact structure of the drug–receptor complex. Br-verapamil is stretched between two subunits of CavAb, consistent with drug binding at the interface of domains III and IV in Cav1.2 channels^{4–6,24,25}. As for dihydropyridines, phenylalkylamine binding at this site disrupts the fourfold symmetry of the pore (Extended Data Fig. 9). Location of the phenylalkylamine receptor site deep in the central cavity in the pore reveals why binding of these drugs is state-dependent. Access of phenylalkylamines to their receptor is greatly enhanced by opening the intracellular activation gate, which allows diffusion to the drug receptor site. Drug binding is therefore frequency-dependent, allowing selective block of Cav1.2 channels in rapidly firing cardiac myocytes². This mechanism is the basis for use of verapamil for cardiac arrhythmias.

Overall, our structural studies illuminate the complex pharmacology and therapeutic uses of Ca²⁺ antagonist drugs in treatment of different cardiovascular disorders at the atomic level (see Supplementary Discussion). These structural models will be important for design and development of next-generation Ca²⁺ antagonist drugs to provide safer and more effective treatment of hypertension, angina pectoris, cardiac arrhythmia, and other medical conditions.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 March; accepted 12 July 2016.

Published online 24 August 2016.

- Zamponi, G. W., Striessnig, J., Koschak, A. & Dolphin, A. C. The physiology, pathology, and pharmacology of voltage-gated calcium channels and their future therapeutic potential. *Pharmacol. Rev.* **67**, 821–870 (2015).
- Hondeghem, L. M. & Katzung, B. G. Antiarrhythmic agents: the modulated receptor mechanism of action of sodium and calcium channel-blocking drugs. *Annu. Rev. Pharmacol. Toxicol.* **24**, 387–423 (1984).
- Murphy, K. M., Gould, R. J., Largent, B. L. & Snyder, S. H. A unitary mechanism of calcium antagonist drug action. *Proc. Natl Acad. Sci. USA* **80**, 860–864 (1983).
- Catterall, W. A. & Striessnig, J. Receptor sites for Ca²⁺ channel antagonists. *Trends Pharmacol. Sci.* **13**, 256–262 (1992).
- Hockerman, G. H., Peterson, B. Z., Johnson, B. D. & Catterall, W. A. Molecular determinants of drug binding and action on L-type calcium channels. *Annu. Rev. Pharmacol. Toxicol.* **37**, 361–396 (1997).
- Striessnig, J. Pharmacology, structure and function of cardiac L-type Ca²⁺ channels. *Cell. Physiol. Biochem.* **9**, 242–269 (1999).
- Hofmann, F., Lacinová, L. & Klugbauer, N. Voltage-dependent calcium channels: from structure to function. *Rev. Physiol. Biochem. Pharmacol.* **139**, 33–87 (1999).
- Cheng, R. C., Tikhonov, D. B. & Zhorov, B. S. Structural model for phenylalkylamine binding to L-type calcium channels. *J. Biol. Chem.* **284**, 28332–28342 (2009).
- Tikhonov, D. B. & Zhorov, B. S. Structural model for dihydropyridine binding to L-type calcium channels. *J. Biol. Chem.* **284**, 19006–19017 (2009).
- Takahashi, M., Seagar, M. J., Jones, J. F., Reber, B. F. & Catterall, W. A. Subunit structure of dihydropyridine-sensitive calcium channels from skeletal muscle. *Proc. Natl Acad. Sci. USA* **84**, 5478–5482 (1987).
- Tanabe, T. et al. Primary structure of the receptor for calcium channel blockers from skeletal muscle. *Nature* **328**, 313–318 (1987).
- Mikami, A. et al. Primary structure and functional expression of the cardiac dihydropyridine-sensitive calcium channel. *Nature* **340**, 230–233 (1989).
- Wu, J. et al. Structure of the voltage-gated calcium channel Cav1.1 complex. *Science* **350**, aad2395 (2015).
- Ren, D. et al. A prokaryotic voltage-gated sodium channel. *Science* **294**, 2372–2375 (2001).
- Catterall, W. A. & Zheng, N. Deciphering voltage-gated Na⁺ and Ca²⁺ channels by studying prokaryotic ancestors. *Trends Biochem. Sci.* **40**, 526–534 (2015).
- Payandeh, J., Scheuer, T., Zheng, N. & Catterall, W. A. The crystal structure of a voltage-gated sodium channel. *Nature* **475**, 353–358 (2011).
- Payandeh, J., Gamal El-Din, T. M., Scheuer, T., Zheng, N. & Catterall, W. A. Crystal structure of a voltage-gated sodium channel in two potentially inactivated states. *Nature* **486**, 135–139 (2012).
- Zhang, X. et al. Crystal structure of an orthologue of the NaChBac voltage-gated sodium channel. *Nature* **486**, 130–134 (2012).
- Tang, L. et al. Structural basis for Ca²⁺ selectivity of a voltage-gated calcium channel. *Nature* **505**, 56–61 (2014).
- Catterall, W. A., Perez-Reyes, E., Snutch, T. P. & Striessnig, J. Voltage-gated calcium channels: introduction. IUPHAR/BPS Guide to Pharmacology (<http://guidetopharmacology.org/GRAC/FamilyIntroductionForward?familyId=80>) (2011).
- Striessnig, J., Murphy, B. J. & Catterall, W. A. Dihydropyridine receptor of L-type Ca²⁺ channels: identification of binding domains for [³H](+)-PN200-110 and [³H]azidopine within the α_1 subunit. *Proc. Natl Acad. Sci. USA* **88**, 10769–10773 (1991).
- Yamaguchi, S. et al. Key roles of Phe1112 and Ser1115 in the pore-forming IIS5-S6 linker of L-type Ca²⁺ channel α_{1C} subunit (Cav 1.2) in binding of dihydropyridines and action of Ca²⁺ channel agonists. *Mol. Pharmacol.* **64**, 235–248 (2003).
- Hescheler, J., Pelzer, D., Trube, G. & Trautwein, W. Does the organic calcium channel blocker D600 act from inside or outside on the cardiac cell membrane? *Pflügers Archiv* **393**, 287–291 (1982).
- Striessnig, J., Glossmann, H. & Catterall, W. A. Identification of a phenylalkylamine binding region within the α_1 subunit of skeletal muscle Ca²⁺ channels. *Proc. Natl Acad. Sci. USA* **87**, 9108–9112 (1990).
- Hockerman, G. H., Johnson, B. D., Scheuer, T. & Catterall, W. A. Molecular determinants of high affinity phenylalkylamine block of L-type calcium channels. *J. Biol. Chem.* **270**, 22119–22122 (1995).
- Yatani, A. & Brown, A. M. The calcium channel blocker nitrendipine blocks sodium channels in neonatal rat cardiac myocytes. *Circ. Res.* **56**, 868–875 (1985).
- Kass, R. S., Arena, J. P. & Chin, S. Block of L-type calcium channels by charged dihydropyridines. Sensitivity to side of application and calcium. *J. Gen. Physiol.* **98**, 63–75 (1991).
- Bangalore, R., Baidur, N., Rutledge, A., Triggie, D. J. & Kass, R. S. L-type calcium channels: asymmetrical intramembrane binding domain revealed by variable length, permanently charged 1,4-dihydropyridines. *Mol. Pharmacol.* **46**, 660–666 (1994).
- Peterson, B. Z. & Catterall, W. A. Calcium binding in the pore of L-type calcium channels modulates high affinity dihydropyridine binding. *J. Biol. Chem.* **270**, 18201–18204 (1995).
- Glossmann, H., Ferry, D. R., Goll, A., Striessnig, J. & Zernig, G. Calcium channels and calcium channel drugs: recent biochemical and biophysical findings. *Arzneimittelforschung* **35**, 1917–1935 (1985).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are grateful to the beamline staff at the Advanced Light Source (BL8.2.1 and BL8.2.2) for their assistance during data collection. Research reported in this publication was supported by the National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health under award number R01 HL112808 (W.A.C. and N.Z.), and a National Research Service Award from training grant T32 GM008268 (T.M.S.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was also supported by research grants from the Howard Hughes Medical Institute (N.Z.) and by the National Institute of Neurological Disorders and Stroke (NINDS) of the National Institutes of Health under award number R01 NS26254 (W.A.C.). D.C.P. acknowledges support from Neusentis, Pfizer Inc., Cambridge, UK during the course of this work.

Author Contributions L.T., T.M.G., T.M.S., T.S., D.C.P., N.Z., and W.A.C. designed the experiments. D.C.P. provided compound UK-59811. L.T. conducted protein purification, crystallization, and X-ray diffraction experiments for dihydropyridines. L.T. and T.M.S. conducted protein purification, crystallization, and X-ray diffraction experiments for Br-verapamil. L.T. and T.M.S. determined the structures and analysed the structural results with input from T.M.G. and N.Z. T.M.G. designed and analysed mutants that block drug binding and performed all of the electrophysiological studies. T.M.G., T.S., and W.A.C. analysed the electrophysiological results. All authors contributed to the interpretation of the structures in light of the physiological data. L.T., N.Z., and W.A.C. wrote the manuscript with input from all co-authors.

Author Information The coordinates and structure factors have been deposited in the Protein Data Bank with the following accession codes 5KLB (CavAb, 5 mM Ca²⁺ 2.7 Å); 5KLG (CavAb-W195Y-UK-59811, 5 mM Ca²⁺); 5KLS (CavAb-UK-59811, 5 mM Ca²⁺); 5KMD (CavAb-W195Y-amlopidine, 5 mM Ca²⁺); 5KMF (CavAb-W195Y-nimodipine, 5 mM Ca²⁺); and 5KMH (CavAb-Br-verapamil, 5 mM Ca²⁺). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.Z. (nzheng@uw.edu) or W.A.C. (wcatt@uw.edu).

METHODS

Ca_vAb constructs and drugs. As originally defined, Ca_vAb was constructed by introducing the mutations E177D, S178D and M181D as a triple mutant into NavAb¹⁹. This construct was used for all electrophysiological studies, except as noted in figure legends. In this work, we have also used Ca_vAb E177D S178D M181N, which has an identical structure and Ca²⁺-binding properties and has high Ca²⁺ selectivity¹⁹. It gives greater consistency of high-resolution crystal structures. We have also added the mutation W195Y, which substitutes the Y residue from the analogous position in mammalian Ca_v1.1 channels for W195 in Ca_vAb. This mutant gives better resolution of drugs bound at the dihydropyridine site. Ca_vAb E177D S178D M181N W195Y was used for all structural studies presented here, except as noted in the figure legends. Similar structural results were obtained for both versions of Ca_vAb. We found that amlodipine and other dihydropyridines (Figs. 1 and 2 and Extended Data Figs 1, 3 and 4) and verapamil and other phenylalkylamines (Fig. 4 and Extended Data Fig. 6) effectively blocked Ca_vAb and gave high-resolution crystal structures; however, we were unable to prepare crystals with bound diltiazem for structural biology so we have not addressed the structure of the benzothiazepine receptor site in this work.

Electrophysiology. All measurements were done in insect cells (*Trichoplusia ni* cells; High5). All Ca_vAb constructs used were made on the background of N49K mutation. Mutation N49K shifts the activation curve ~75 mV to more positive potentials compared to wild-type Ca_vAb and abolishes the use-dependent inactivation as described previously^{19,31}. All constructs showed good expression, allowing measurement of ionic currents 24–48 h after infection. Whole-cell Ba²⁺ currents were recorded using an Axopatch 200 amplifier (Molecular Devices) with glass micropipettes (2–4 MΩ). Capacitance was subtracted and 80–90% of series resistance was compensated using internal amplifier circuitry. Extracellular solution contained in (mM) 10 BaCl₂, 140 NMDG-methanesulphonate, 20 HEPES, (pH 7.4, adjusted with Ba(OH)₂, [Ba²⁺]_{total} = 13 mM). Intracellular solution contained in (mM) 105 CsF, 35 NaCl, 10 HEPES, 10 EGTA, (pH 7.4, adjusted with CsOH). Current–voltage (*I*–*V*) relationships were recorded in response to steps to voltages ranging from –120 to +50 mV in 10-mV increments from a holding potential of –120 mV. Conductance–voltage (*G*–*V*) curves were calculated from the corresponding (*I*–*V*) curves. Pulses were generated and currents were recorded using Pulse software controlling an Instrutech ITC18 interface (HEKA). Data were analysed using Igor Pro 6.2 (WaveMetrics). Sample sizes were chosen to give s.e.m. values of less than 10% of peak values based on prior experimental experience. Inhibition curves were fit with a Hill equation with *n*_H = 1.0 unless indicated otherwise in the figure legends.

Protein expression and purification. The pFastBac-Flag-Ca_vAb was used as the construct for producing homotetrameric model voltage-gated Ca²⁺ channel¹⁹. I199S, W195Y, and T206S constructs were generated via site-directed mutagenesis using QuickChange (Stratagene). Recombinant baculovirus were produced using the Bac-to-Bac system (Invitrogen), and *T. ni* insect cells were infected for large-scale protein purification. Cells were harvested 72 h post-infection and resuspended in buffer A (50 mM Tris-HCl, pH = 8.0, 200 mM NaCl) supplemented with protease inhibitors and DNase. After sonication, digitonin (EMD Biosciences) was added to 1%, and solubilization was carried out for 1–2 h at 4 °C. Clarified supernatant was then incubated with anti-Flag M2-agarose resin (Sigma) for 1–2 h at 4 °C with gentle mixing. Flag-resin was washed with ten column volumes of buffer B (buffer A supplemented with 0.12% digitonin) and eluted with buffer B supplemented with 0.1 mg ml^{–1} Flag peptide. The eluant was concentrated and then passed over a Superdex 200 column (GE Healthcare) in 10 mM Tris-HCl pH = 8.0, 100 mM NaCl and 0.12% digitonin. The peak fractions were concentrated using a Vivaspinn 30K centrifugal device.

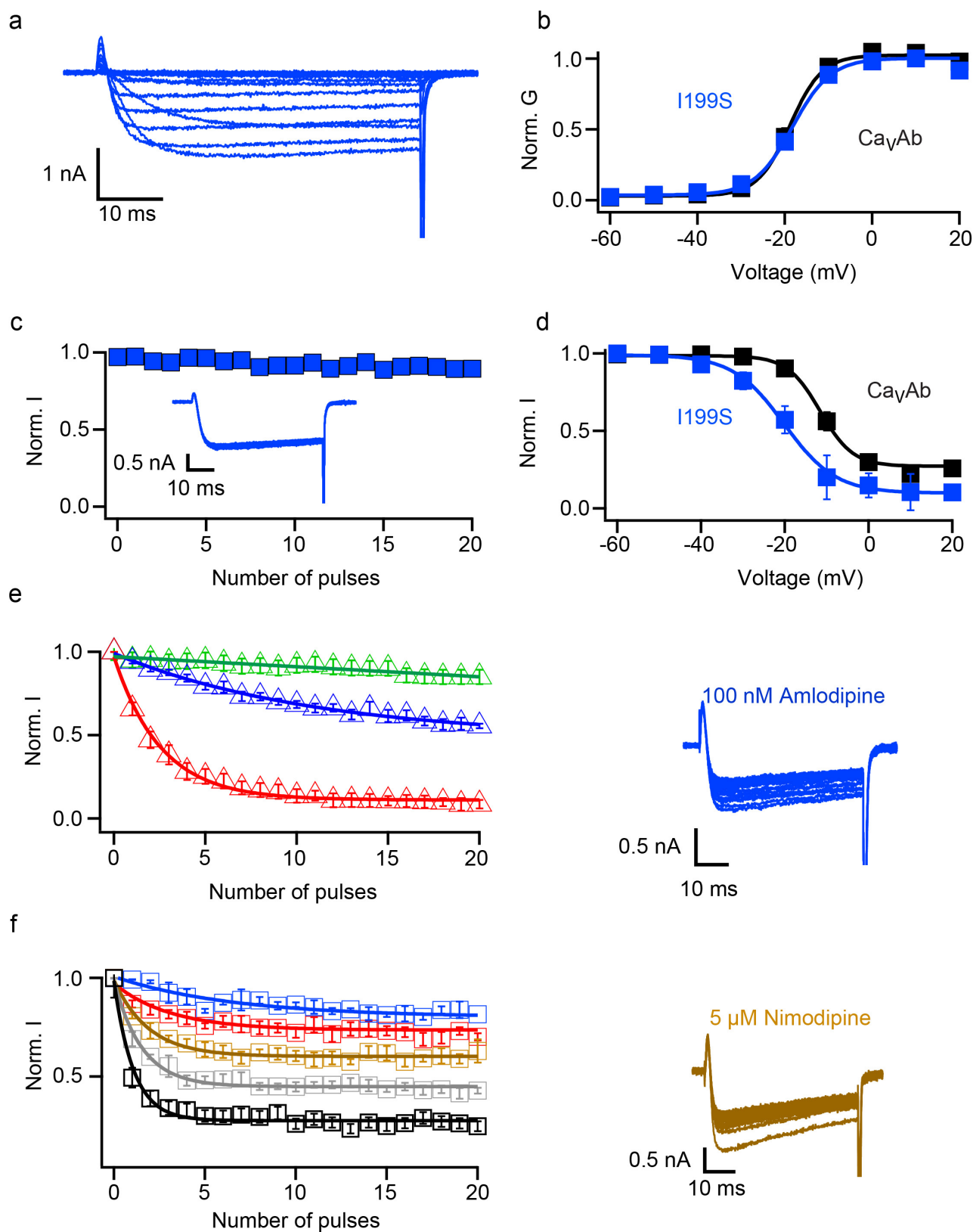
Crystallization and data collection. Ca_vAb and the W195Y mutant were concentrated to ~20 mg ml^{–1} and reconstituted into DMPC:CHAPSO (Anatrace) bicelles

according to standard protocols^{16,17,32,33}. The protein-bicelle preparation and a well solution containing 1.8–2.0 M ammonium sulphate, 100 mM Na-citrate pH = 5.0 was mixed in 1:1 ratio and set up in a hanging-drop vapour-diffusion format. All the antagonist complex crystals were obtained through co-crystallization by incubating the protein-bicelle with 100 μM antagonist overnight before setting up crystallization trials. For the UK-59811 complex, both 100 μM and 200 μM antagonist were used for UK-59811/Ca_vAb crystals. Crystals were cryoprotected by soaking in 0.1 M Na-acetate, pH 5.0, 26% glucose, 2.0 M ammonium sulphate, and 5 mM Ca²⁺. Crystals were plunged into liquid nitrogen and maintained at 100 K during all data collection procedures.

The anomalous diffraction data sets for Br were collected at 0.9194 Å, and the anomalous data sets for Ca²⁺ were collected at 1.75 Å with the same synchrotron radiation source (Advanced Light Source, BL8.2.1). To optimize the anomalous scattering signal, the data sets were collected by using the ‘inverse beam strategy’ with the wedge size of 5°.

Structure determination, refinement, and analysis. X-ray diffraction data were integrated and scaled with the HKL2000 package³⁴ and further processed with the CCP4 package³⁵. The structure of Ca_vAb and its antagonist complex were solved by molecular replacement using an individual subunit of the Ca_vAb structure (PDB code 4MS2) as the search template. The data sets were processed in P21221 space group, in which there are four molecules in one asymmetric unit. Crystallography and NMR System software³⁶ were used for refinement of coordinates and *B*-factors. Final models were obtained after several cycles of refinement with REFMAC³⁷ and PHENIX³⁸ plus manual re-building using COOT³⁹. The geometries of the final structural models of Ca_vAb and its antagonist complexes were verified using PROCHECK⁴⁰. Divalent cations were identified by anomalous difference Fourier maps calculated using data collected at wavelengths of 1.75 Å for Ca²⁺. The Br atoms of UK-59811 and Br-verapamil were identified by anomalous difference Fourier maps calculated using data collected at wavelengths of 0.9194 Å. Procedures accounting for merohedral twinning were performed during structural refinement of amlodipine, nimodipine, and Br-verapamil data sets. Detailed crystallographic data and refinement statistics for all constructs are shown in Extended Data Table 1. All structural figures were prepared with PyMol⁴¹.

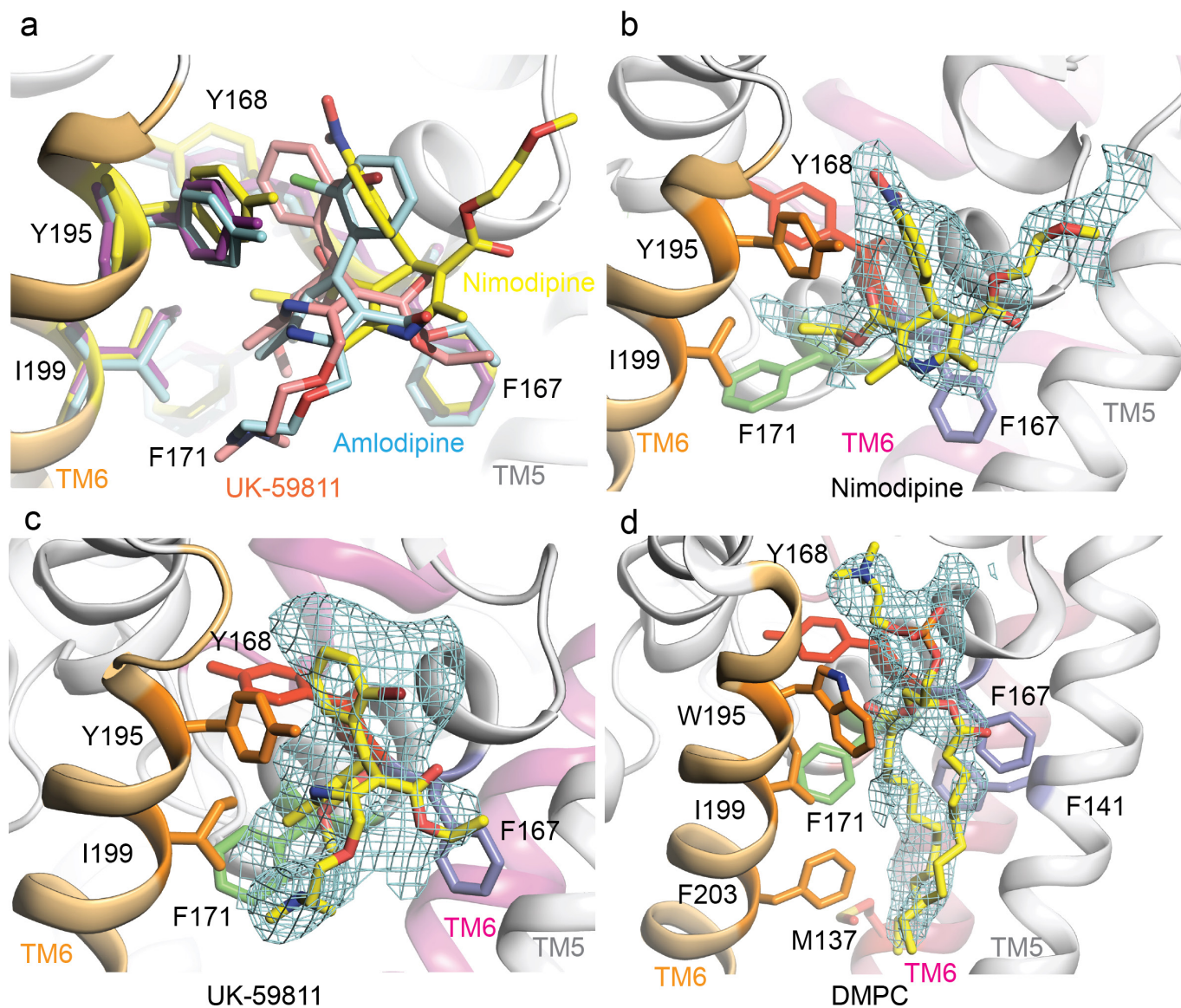
31. Gamal El-Din, T. M., Martinez, G. Q., Payandeh, J., Scheuer, T. & Catterall, W. A. A gating charge interaction required for late slow inactivation of the bacterial sodium channel NavAb. *J. Gen. Physiol.* **142**, 181–190 (2013).
32. Faham, S. & Bowie, J. U. Bicelle crystallization: a new method for crystallizing membrane proteins yields a monomeric bacteriorhodopsin structure. *J. Mol. Biol.* **316**, 1–6 (2002).
33. Faham, S. *et al.* Crystallization of bacteriorhodopsin from bicelle formulations at room temperature. *Protein Sci.* **14**, 836–840 (2005).
34. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
35. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
36. Brünger, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
37. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
38. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
39. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
40. Laskowski, R. A., Moss, D. S. & Thornton, J. M. Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.* **231**, 1049–1067 (1993).
41. DeLano, W. L. PyMol molecular viewer (V1.2r3pre) (<http://www.pymol.org>) (2002).



Extended Data Figure 1 | Biophysical characterization of Ca_vAb I199S.

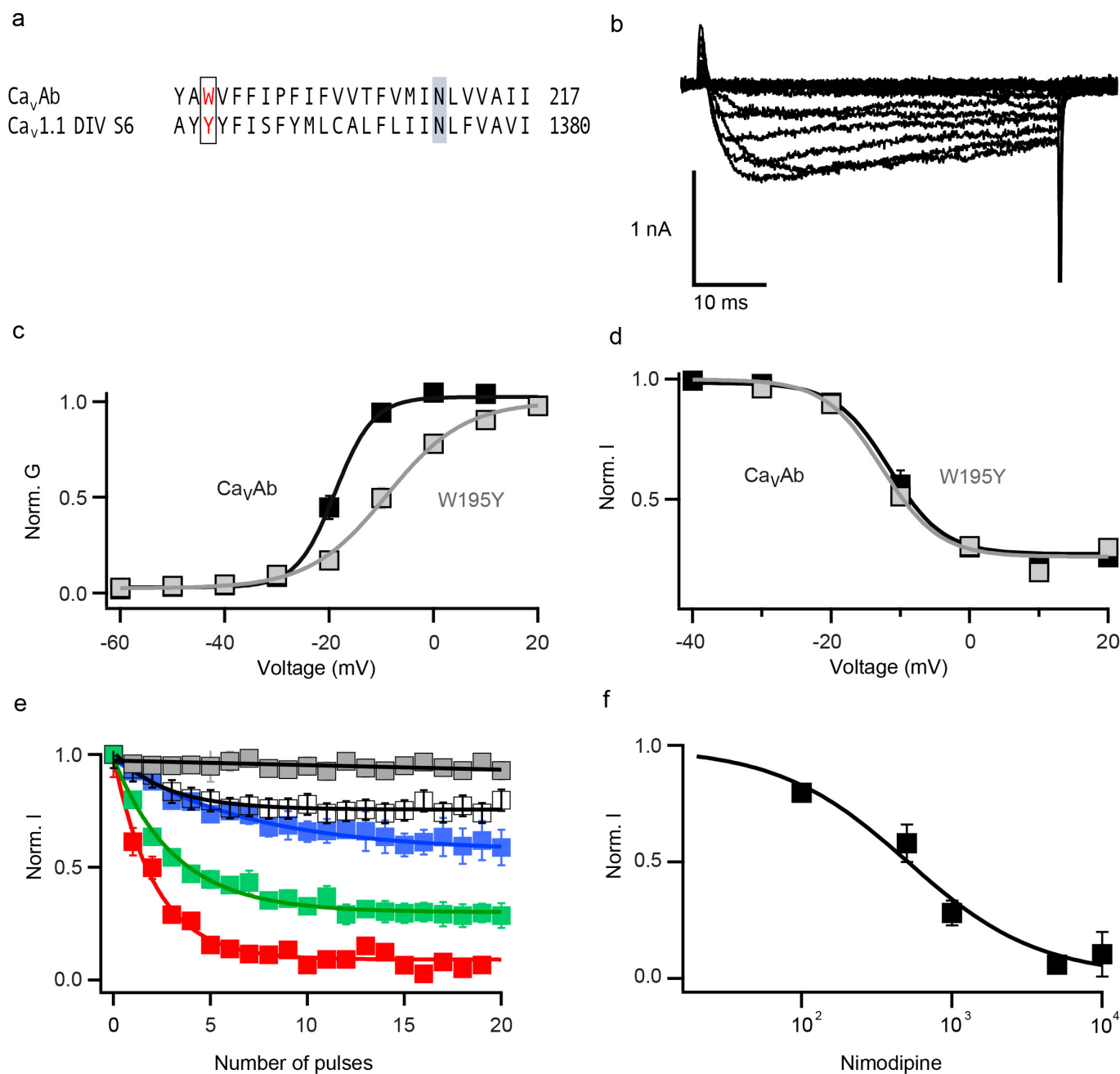
a, Ba^{2+} currents recorded from a holding potential of -120 mV to test potentials from -60 mV to 20 mV in 10 mV steps for I199S. **b**, G - V curves of Ca_vAb and Ca_vAb I199S derived from peak I - V relationships. The voltages for half-maximal activation and slopes are: Ca_vAb : $V_{1/2} = -18.8 \pm 0.3$, $k = 3.68 \pm 0.43$, $n = 7$; Ca_vAb I199S: $V_{1/2} = -18.8 \pm 0.3$, $k = 3.88 \pm 0.47$ ($n = 5$). **c**, Repetitive depolarization to 0 mV at 1 Hz from a holding potential of -120 mV ($n = 5$). **d**, Steady-state inactivation of Ca_vAb and Ca_vAb I199S. Two pulses were applied: a 300 -ms conditioning pulse to the indicated

potentials followed by 50 -ms test pulse to 0 mV ($n = 3$). **e**, State-dependent block of Ca_vAb I199S by 10 nM (green), 100 nM (blue), or 1.5 μM (red) amlodipine during repetitive depolarizations to 0 mV (left, $n = 3$ – 5 cells). Ba^{2+} currents in 100 nM amlodipine for Ca_vAb I199S (right). **f**, Concentration-dependent block of Ca_vAb I199S by nimodipine at 100 nM (blue), 1 μM (red), 5 μM (brown), 10 μM (grey) and 50 μM (black) (left, $n = 4$ – 5 cells for each curve). Ba^{2+} currents in the presence of 5 μM nimodipine for Ca_vAb I199S (right).



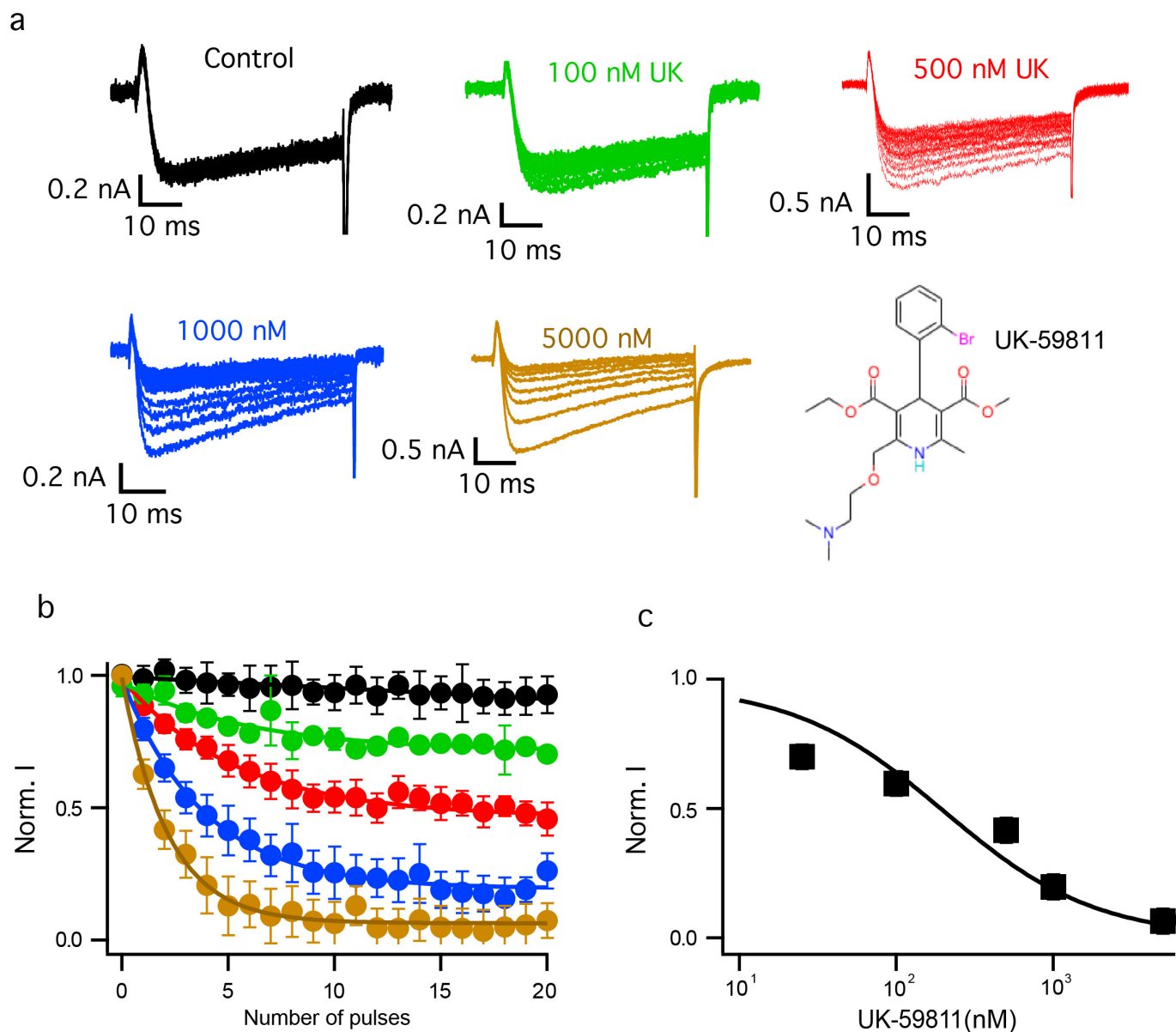
Extended Data Figure 2 | Structural comparison of the binding modes of amlodipine, nimodipine, and UK-59811. **a**, Superposition of CavAb in complexes with amlodipine (cyan), nimodipine (yellow), and UK-59811 (magenta) at the dihydropyridine binding site viewed from the side of the pore module. The side chains of dihydropyridine-interacting residues are

shown in sticks. **b**, An *Fo*–*Fc* simulated annealing omit map contoured at 2.5σ for nimodipine. **c**, An *Fo*–*Fc* simulated annealing omit map contoured at 2.5σ for UK-59811. **d**, An *Fo*–*Fc* simulated annealing omit map contoured at 2.5σ for DMPC.



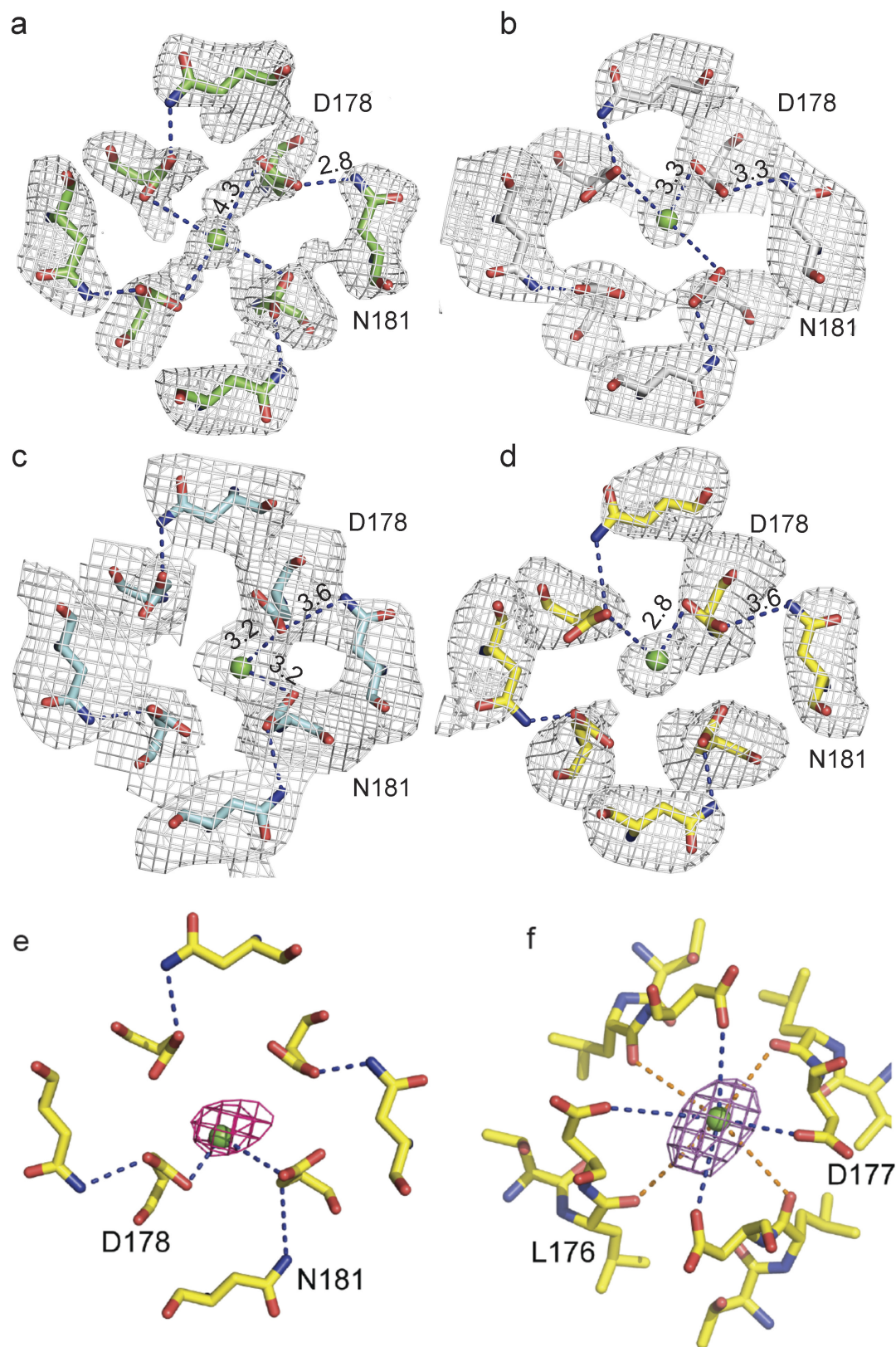
Extended Data Figure 3 | Biophysical characterization and drug block of Ca_vAb W195 and Ca_vAb Y195. **a**, Sequence alignment of Ca_vAb S6 segment and Ca_v1.1 DIV S6. W195 in Ca_vAb is equivalent to Y1358 in Ca_v1.1. **b**, Ba²⁺ currents recorded from a holding potential of -120 mV to test potentials from -60 mV to 20 mV in 10 mV steps for Ca_vAb W195Y. **c**, *G*-*V* curves for Ca_vAb W195 and Ca_vAb Y195 derived from peak *I*-*V* relationships. The voltages for half-maximal activation and slopes are: Ca_vAb W195 $V_{1/2} = -18.8 \pm 0.3$, $k = 3.7 \pm 0.43$, $n = 7$; Ca_vAb Y195,

$V_{1/2} = -9 \pm 0.3$, $k = 7.4 \pm 0.1$, $n = 5$. **d**, Steady-state inactivation of Ca_vAb W195 and Ca_vAb Y195 ($n = 3$). Two pulses were applied: a 300-ms conditioning pulse followed by 50-ms test pulse to 0 mV. **e**, State-dependent block of Ca_vAb W195Y by nimodipine at 100 nM (white), 500 nM (blue), 1 μ M (green), 5 μ M (red), and control (grey). **f**, Concentration-dependent block of Ca_vAb W195Y by nimodipine. $IC_{50} = 508 \pm 93$ nM ($n = 4-5$ cells for each point).



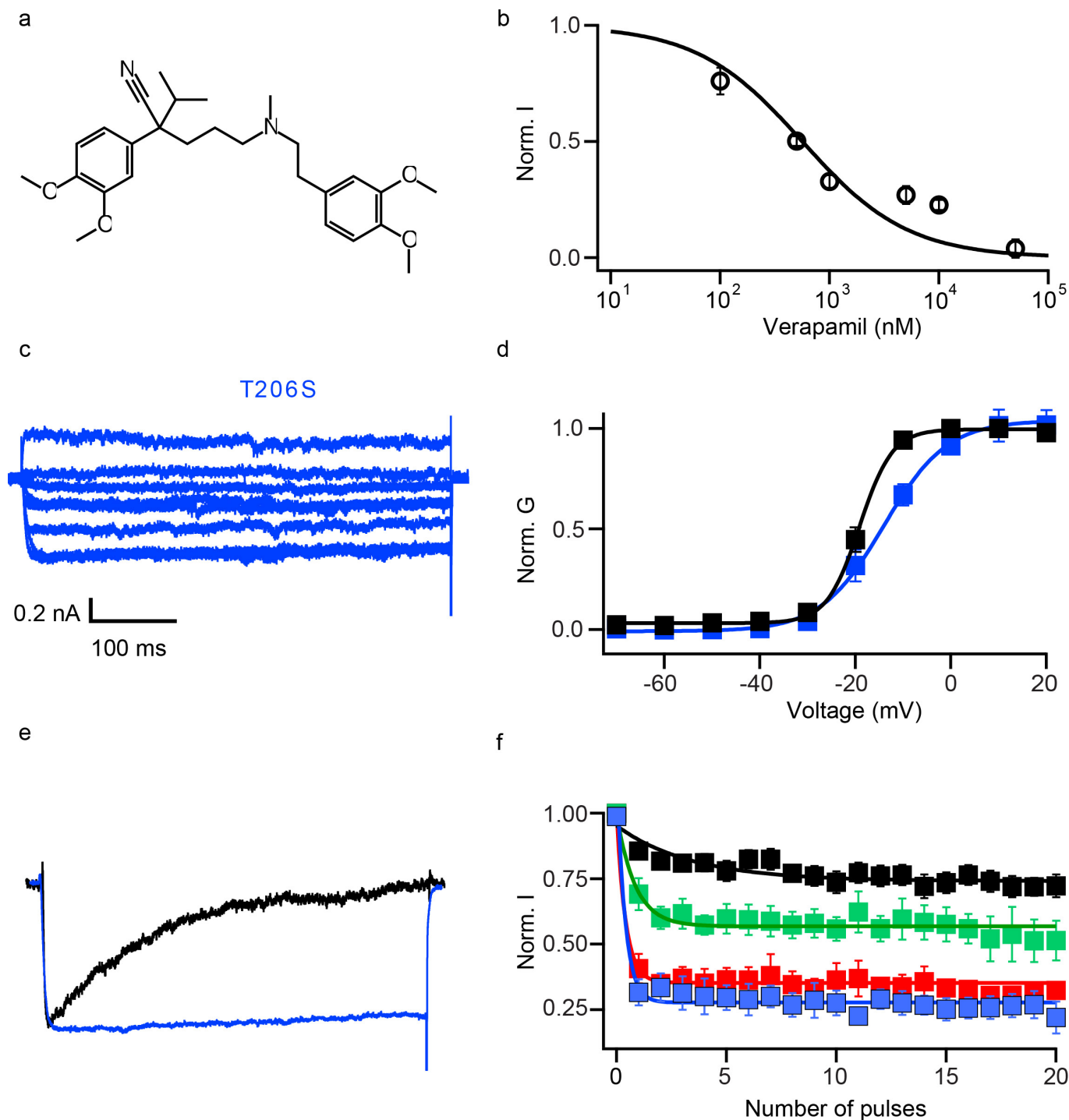
Extended Data Figure 4 | Biophysical characterization of block by UK-59811. **a**, Ba^{2+} currents for state-dependent block by different concentrations of UK-59811. **b**, State-dependent block of Ca_vAb by UK-59811 at 0 nM (black), 100 nM (green), 500 nM (red), 1 μM (blue),

and 5 μM (brown). For each curve, $n = 4-5$ cells. **c**, Concentration-response curve for UK-59811. Data were fit with a Hill equation assuming a 1:1 binding. $\text{IC}_{50} = 194 \pm 22$ nM, $n = 4-5$.



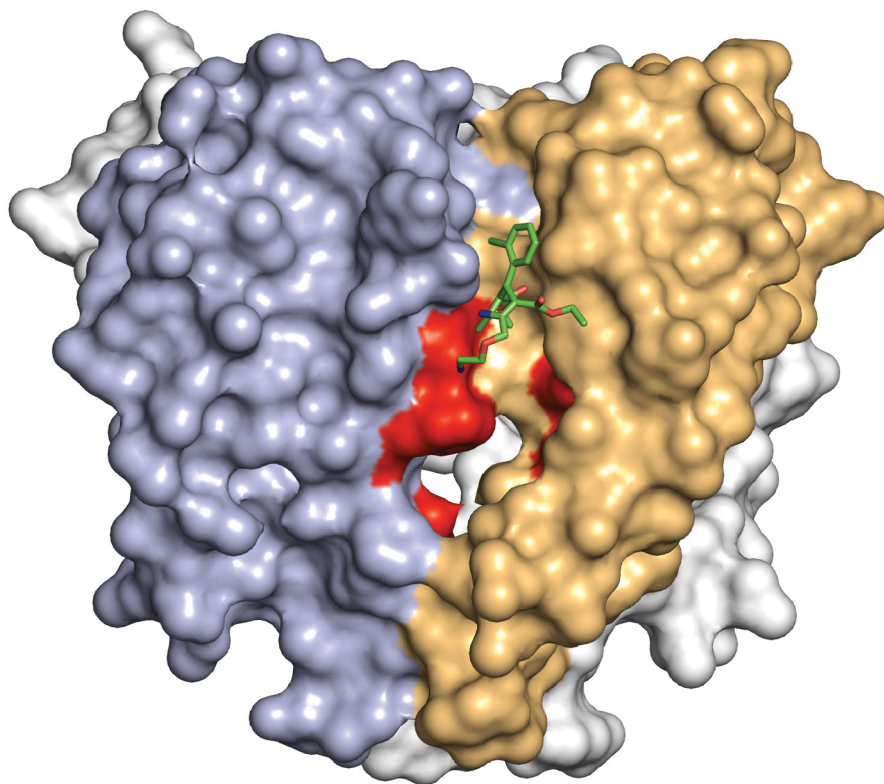
Extended Data Figure 5 | Evidence for the partially dehydrated Ca^{2+} binding and carboxyl-carboxylate pairs at the selectivity filter entryway. **a**, Top view of an *Fo*–*Fc* simulated annealing omit map contoured at 3σ for residues 178 and 181 for the wild-type channel without drug. **b**, Top view of an *Fo*–*Fc* simulated annealing omit map contoured at 3σ for residues 178 and 181 for *Ca_vAb*–amlodipine. **c**, Top view of an *Fo*–*Fc* simulated annealing omit map contoured at 2.5σ for residues 178 and 181 for *Ca_vAb*–nimodipine. **d**, Top view of an *Fo*–*Fc* simulated

annealing omit map contoured at 3σ for residues 178 and 181 for *Ca_vAb*–UK-59811. **e**, Top view of Site 1 with the anomalous difference Fourier map density (red mesh, contoured at 3σ) calculated with diffraction data of crystals collected at 1.75 Å wavelength. Ca^{2+} is shown as a green sphere. Site 1 residues are shown in sticks. Hydrogen bonds are indicated with dashed lines. **f**, Top view of Site 2 with the anomalous difference Fourier map density (magenta mesh, contoured at 3σ).



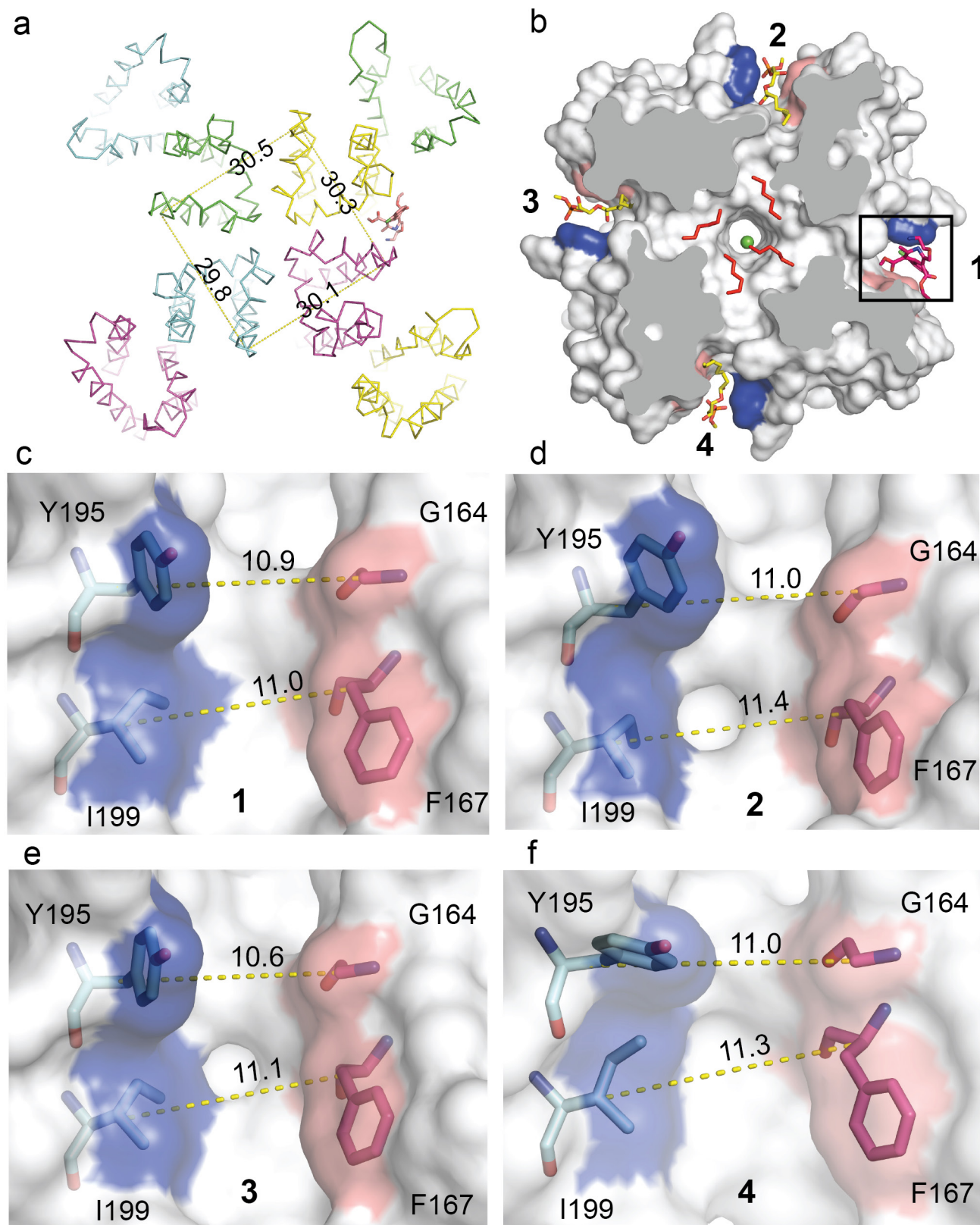
Extended Data Figure 6 | Biophysical characterization of verapamil block of $\text{Ca}_\text{V}\text{Ab}$ and functional properties of $\text{Ca}_\text{V}\text{Ab}$ T206S. **a**, Chemical structure of verapamil. **b**, Concentration dependence of verapamil inhibition of $\text{Ca}_\text{V}\text{Ab}$. The amplitude of the peak Ba^{2+} current was recorded after applying 20 pulses at a frequency of 1 Hz, where the block reaches steady state. The data were fit by a Hill equation assuming a 1:1 binding ratio. $n = 4-7$ cells. $\text{IC}_{50} = 475 \pm 25$ nM. **c**, Ba^{2+} currents of $\text{Ca}_\text{V}\text{Ab}$ T206S.

d, $G-V$ curves. $\text{Ca}_\text{V}\text{Ab}$ (black): $V_{1/2} = 18.8 \pm 0.3$ mV, $k = 3.7 \pm 0.43$ ($n = 5$); $\text{Ca}_\text{V}\text{Ab}$ T206S (blue): $V_{1/2} = -15 \pm 1.8$ mV, $k = 6.6 \pm 0.4$ ($n = 5$). **e**, Current traces of $\text{Ca}_\text{V}\text{Ab}$ (black) and $\text{Ca}_\text{V}\text{Ab}$ T206S (blue) during a 1-s depolarizing pulse from a holding potential of -120 mV to -10 mV. **f**, State-dependent inhibition of $\text{Ca}_\text{V}\text{Ab}$ T206S by Br-verapamil at $10 \mu\text{M}$ (black), $25 \mu\text{M}$ (green), $50 \mu\text{M}$ (red), and $100 \mu\text{M}$ (blue). For each curve, $n = 4-5$ cells.



Extended Data Figure 7 | Comparison of dihydropyridine binding site in Ca_vAb and Ca_v1.2. The pore domain of Ca_vAb is illustrated with two subunits in view, one in tan corresponding to domain III of Ca_v1.2 and one in blue corresponding to domain IV of Ca_v1.2. The amino

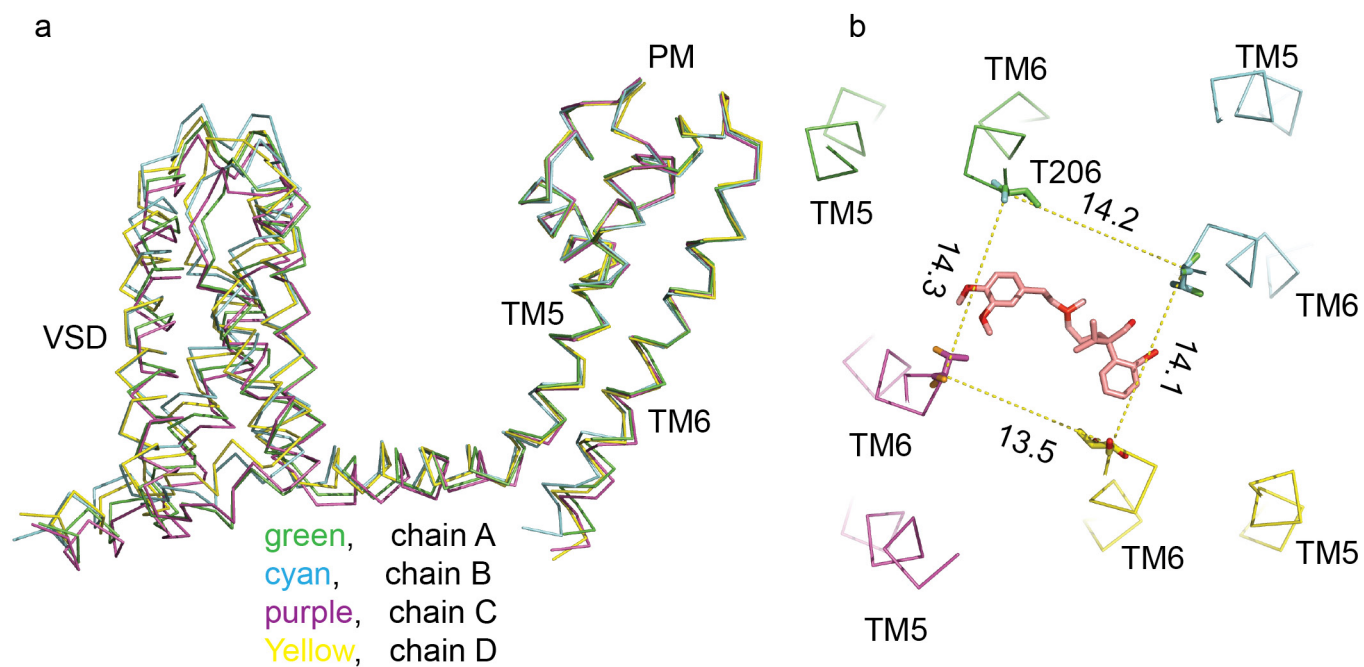
acid residues in Ca_vAb corresponding to those that are important for dihydropyridine binding to Ca_v1.2 channels are highlighted in red. Bound amlodipine is illustrated with green sticks.



Extended Data Figure 8 | Amlodipine binding breaks symmetry.

a, The overall structure of CavAb in complex with amlodipine (shown in ribbon representation). Measuring the C_{α} distances of V196 (nearing the amlodipine binding pocket) from the 4 subunits shows the channel is asymmetrical. **b**, Binding of amlodipine (sticks in red) induces asymmetry and causes rearrangement of the lipid in the central cavity. **c–f**, The amlodipine binding pocket showing the C_{α} – C_{α} distance at two layers

(Y195–G164 and I199–F167) horizontally. At layer 1 (Y195–G164), the C_{α} – C_{α} distance of its neighbouring sites (11.0 Å in **d** and 11.0 Å in **f**) matches the drug binding site (10.9 Å in **c**), but the diagonal site (**e**) is too narrow (10.6 Å). At layer 2 (I199–F167), the pocket width of the diagonal site (11.1 Å in **e**) matches the drug-binding site (11.0 Å in **c**), but the two diagonal sites are too wide (11.4 Å in **d** and 11.3 Å in **f**).



Extended Data Figure 9 | Br-verapamil binding breaks symmetry. **a**, Alignment of the 4 subunits of CavAb in complex with Br-verapamil showing the voltage sensor module (VSD) and the ends of S6 are different. **b**, Measuring the C_{α} distances between T206 residues in adjacent subunits shows that the channel is indeed asymmetrical with Br-verapamil in the pore.

Extended Data Table 1 | Data collection, phasing and refinement statistics

	Ca _v Ab 5mM Ca ²⁺	Ca _v Ab ¹ (W195Y) UK-59811 5mM Ca ²⁺	Ca _v Ab ² UK-59811 5mM Ca ²⁺	Ca _v Ab (W195Y) Amlodipine 5mM Ca ²⁺	Ca _v Ab ¹ (W195Y) Nimodipine 5mM Ca ²⁺	Ca _v Ab ¹ Br-verapamil 5mM Ca ²⁺
Data collection						
Space group	P21221	P21221	P21221	P21221	P21221	P21221
Cell dimensions						
<i>a</i> , <i>b</i> , <i>c</i> (Å)	124.9 125.7 191.5	125.9 126.0 192.1	125.5 125.9 191.7	125.6 125.3 191.7	125.3 125.4 191.6	125.6 125.6 192
α , β , γ (°)	90 90 90	90 90 90	90 90 90	90 90 90	90 90 90	90 90 90
Resolution (Å)	2.7	3.3	3.3	3.2	3.2	3.3
<i>R</i> _{sym} or <i>R</i> _{merge}	11.4(98.4)	12.6(74.2)	11.7(60.1)	12.1(49.1)	11.2(68.6)	18.8(86.1)
CC _{1/2} (%)	99.8(87.7)	99.8(84.3)	99.7(87.4)	99.4(89.2)	99.7(86.7)	98.4(70.3)
<i>I</i> / <i>s</i>	13.4(2.4)	13.4(3.3)	13.1(3.1)	10.2(2.8)	15.2(3.5)	6.0(1.7)
Completeness (%)	92.5(97.8)	95.0(100.0)	92.1(82.0)	92.7(94.5)	99.8(99.8)	97.7(98.8)
Redundancy	10.1(9.8)	9.5(9.9)	9.4(9.2)	5.1(5.2)	9.3(9.0)	4.9(5.0)
Refinement						
Resolution (Å)	30-2.7	30-3.3	30-3.3	30-3.2	30-3.2	30-3.2
No. reflections	76513	46606	42515	46657	50390	49327
<i>R</i> _{work} / <i>R</i> _{free}	22.1/26.2	28.0/30.5	27.5/30.0	23.3/27.7	21.7/25.6	25.1/29.4
No. atoms	9684	7400	7403	7380	7393	7366
Protein	8780	7192	7200	7192	7192	7200
Ligand/ion	887	205	199	187	189	166
Water	17	3	2	1	28	
B-factors						
Protein	76.9	104.4	111.4	100.9	108.0	114.1
Ligand/ion	74.2	99.9	95.8	85.7	86.3	100.6
Water	46.6	57.9	63.2	48.8	52.8	
R.m.s deviations						
Bond lengths (Å)	0.009	0.013	0.013	0.013	0.013	0.013
Bond angles (°)	1.15	1.74	1.75	1.73	1.54	1.74
Ramachandran statistics						
Favored	96%	96%	96%	95.0%	96%	96.0%
Allowed	4.1%	3.6%	3.6%	4.6%	4.0%	4.1%
Outliers	0.28%	0.23%	0.23%	0.23%	0.12%	0.12%

¹This data set is collected at 0.9198 Å.²This data set is collected at 1.75 Å.

All other data sets are collected at 1.0 Å.

CORRIGENDUM

doi:10.1038/nature18280

Corrigendum: A novel multiple-stage antimalarial agent that inhibits protein synthesis

Beatriz Baragaña, Irene Hallyburton, Marcus C. S. Lee, Neil R. Norcross, Raffaella Grimaldi, Thomas D. Otto, William R. Proto, Andrew M. Blagborough, Stephan Meister, Grennady Wirjanata, Andrea Ruecker, Leanna M. Upton, Tara S. Abraham, Mariana J. Almeida, Anupam Pradhan, Achim Porzelle, María Santos Martínez, Judith M. Bolscher, Andrew Woodland, Torsten Luksch¹, Suzanne Norval, Fabio Zuccotto, John Thomas, Frederick Simeons, Laste Stojanovski, Maria Osuna-Cabello, Paddy M. Brock, Tom S. Churcher, Katarzyna A. Sala, Sara E. Zakutansky, María Belén Jiménez-Díaz, Laura Maria Sanz, Jennifer Riley, Rajshekhar Basak, Michael Campbell, Vicky M. Avery, Robert W. Sauerwein, Koen J. Dechering, Rintis Noviyanti, Brice Campo, Julie A. Frearson, Iñigo Angulo-Barturen, Santiago Ferrer-Bazaga, Francisco Javier Gamo, Paul G. Wyatt, Didier Leroy, Peter Siegl, Michael J. Delves, Dennis E. Kyle, Sergio Wittlin, Jutta Marfurt, Ric N. Price, Robert E. Sinden, Elizabeth A. Winzeler, Susan A. Charman, Lidiya Bebrevska, David W. Gray, Simon Campbell, Alan H. Fairlamb, Paul A. Willis, Julian C. Rayner, David A. Fidock, Kevin D. Read & Ian H. Gilbert

Nature **522**, 315–320 (2015); doi:10.1038/nature14451

In this Article, Torsten Luksch was inadvertently omitted from the author list. He is affiliated with the Drug Discovery Unit, Division of Biological Chemistry and Drug Discovery, College of Life Sciences, University of Dundee, Dundee DD1 5EH, UK. His contribution was the analysis of initial screening data alongside author A.W. The online versions of the paper have been corrected.

CORRIGENDUM

doi:10.1038/nature18623

Corrigendum: Robust neuronal dynamics in premotor cortex during motor planning

Nuo Li, Kayvon Daie, Karel Svoboda & Shaul Druckmann

Nature **532**, 459–464 (2016); doi:10.1038/nature17643

We would like to correct several minor errors in this Article. In the Fig. 2b legend, ‘ $*P < 0.01$ ’ should have read ‘ $**P < 0.01$ ’. In the Methods ‘Photoinhibition’ section, the description of galvo step time during photoinhibition of multiple cortical locations should have read ‘step time: <0.2 ms; dwell time: >4.8 ms’ instead of ‘step time: <4.8 ms;’. In Extended Data Fig. 4c, the y -axis values of the three bottom panels should have run from -1 to 1 , instead of from 0 to 2 . In the Extended Data Fig. 5a legend, the citation given for Tlx_PL56-Cre mice should have been to ref. 46, rather than ref. 50. In the Extended Data Fig. 6b legend, the sessions were incorrectly referred to as ‘lick-right trials (session 1, 4)’ and ‘control lick-right trajectories (session 2, 3, 5)’ instead of ‘lick-right trials (session 1, 3, 4)’ and ‘lick-right trajectories (session 2, 5)’. All of these errors have been corrected online, and none of them affects the description, interpretation or conclusions of the Article.

CORRECTIONS & AMENDMENTS

CORRIGENDUM

doi:10.1038/nature18937

Corrigendum: Convection in a volatile nitrogen–ice–rich layer drives Pluto’s geological vigour

William B. McKinnon, Francis Nimmo, Teresa Wong, Paul M. Schenk, Oliver L. White, J. H. Roberts, J. M. Moore, J. R. Spencer, A. D. Howard, O. M. Umurhan, S. A. Stern, H. A. Weaver, C. B. Olkin, L. A. Young, K. E. Smith & the New Horizons Geology, Geophysics and Imaging Theme Team

Nature **534**, 82–85 (2016); doi:10.1038/nature18289

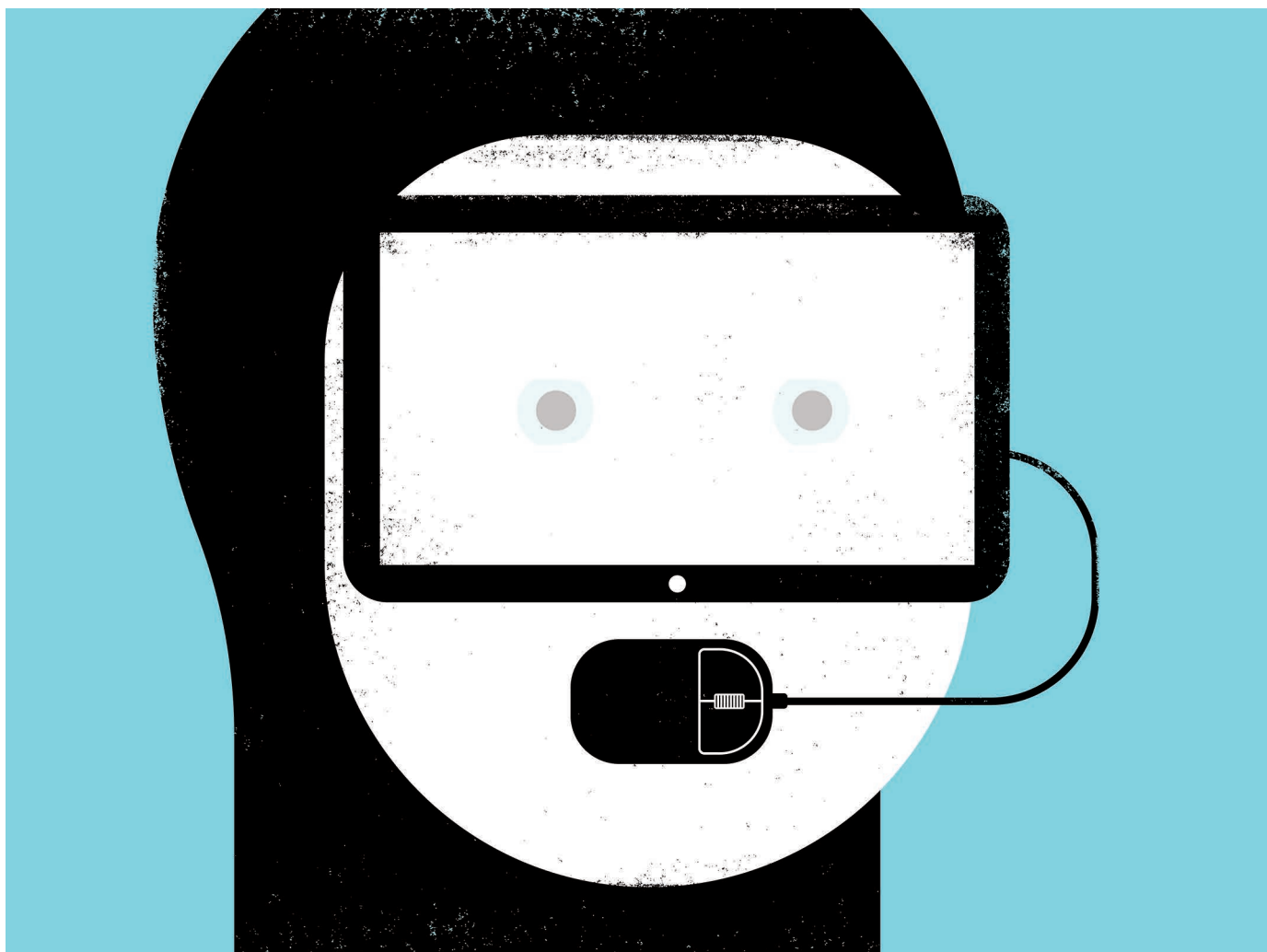
In the list of the New Horizons Geology, Geophysics and Imaging Theme Team, two members were inadvertently omitted: Richard P. Binzel and Alissa Earle (both affiliated with Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA). In addition, the vertical scale in Fig. 3 should be read in metres, not kilometres. These errors have been corrected in the online versions of the Letter.

TOOLBOX

COMPUTERS ON THE REEF

Tools that analyse underwater images of the world's coral reefs are transforming marine ecology.

THE PROJECT TWINS



BY JEFF TOLLEFSON

Blue spires seem to pop out of the photograph in one place; a patch of bushy forms of pinkish-purple in another. To the untrained eye, each is distinct and clearly a coral. Then, Manuel González-Rivero points to a third cluster. The bulbous shapes look like coral, except the smooth grey surface isn't quite right. "The texture

and colour suggest that you are probably not looking at a coral," he says. "More likely you are looking at a crustose coralline alga."

The image is a high-resolution panoramic photograph collected by the XL Catlin Seaview Survey, a scientific initiative that began to catalogue the world's reefs in 2012. To understand how coral reefs are responding to overfishing, pollution, global warming and ocean acidification, the Catlin team — ecologist

González-Rivero among them — is documenting coral abundance, health, structure and biodiversity in millions of underwater snapshots.

It would take decades to go through all these images manually, even with González-Rivero's expert eyes. But the Catlin team is using a neural-networking algorithm: a deep-learning system in which a computer learns to classify what it sees in coral-reef pictures. The project was led by computer scientist ►

► Oscar Beijbom at the University of California, Berkeley, and the software can zip through Catlin's gigantic photo album — currently around one million photographs — in a matter of months.

The software is just one example of how coral researchers are embracing advances in computer science and software to speed up under-sea mapping of reefs around the world. Combined with high-quality imagery and sensors that collect standardized biological data about reefs, these tools could unleash an era of semi-automated data collection and monitoring, freeing up ecologists to spend less time processing data and more time doing research.

"It's a tremendous step forward," says Mark Eakin, who manages the Coral Reef Watch programme for the US National Oceanic and Atmospheric Administration (NOAA) in College Park, Maryland. "When you aren't limited by the speed of people going through and manually processing images, the yield of information is just so much greater."

OCEAN OF DATA

Coral researchers' entry into the world of big data comes none too soon. Long limited by the size of their diving fins and the capacity of their oxygen tanks, marine ecologists are racing to expand their surveys to document and understand the longer-term impacts of rising ocean temperatures and acidification. The bleaching of corals around the world that has accompanied the epic 2015–16 El Niño warming event in the tropical Pacific Ocean has only heightened concerns.

González-Rivero's goal is to cover as much territory as possible to get a sense of how different corals and reefs are responding to these stresses. Computers will never replace the human eye, nor will they obviate the need for detailed underwater investigations and laboratory research, but they can speed up the basic surveys, he says. "What we are trying to do is find a compromise where we get enough information to understand the reef, but at a much faster pace and in a much cheaper way."

The quality does not have to be compromised: according to Beijbom's unpublished results, the deep-learning system agrees with the human eye on features in coral photos about 81% of the time — impressive considering that even two experts are likely to agree only 84% of the time.

Beijbom plans to launch the algorithm in a few months' time for anyone submitting pictures to his website CoralNet, which already uses computer-assisted systems to help the automated analysis of images. The service is free thanks to funding from NOAA, and 420 users from a variety of institutions, including NOAA, have already uploaded nearly 269,000 images to the site. The best results seem to come from use of a semi-automated program in which the computer does simple analyses and alerts human experts to cases that

it's not confident about, Beijbom says.

In many ways, González-Rivero says, marine science is catching up with the terrestrial sciences, which have been developing tools to gather and process copious amounts of data from satellites and aircraft for decades. The software and hardware can't be directly translated to analysing seas, however: the ocean swallows light, so it is difficult to study anything but the shallowest reefs from above.

"We need an army of people making high-quality measurements."

That has pushed coral researchers to adapt the tools. At Michigan State University in East Lansing, for example, biophysicists David Kramer and Atsuko Kanazawa have modified a handheld sensor originally designed for agricultural research.

When used on land, the sensor measures information such as fluorescence in plants, the carbon content of the soil, the temperature of the air and the humidity. Around 300 sensors are in use in 18 countries, and every time a researcher or a government official takes a reading, the data are uploaded to a central server for analysis.

The modified system, dubbed CoralspeQ, pings reefs with different kinds of light and records the returning spectral signal in 256 wavelengths, from ultraviolet to infrared. These data can be used to measure a reef's photosynthetic activity, for instance, by measuring the fluorescence of chlorophyll in symbiotic algae that provide their host corals with oxygen and nutrients. Knowing how much photosynthetic activity is taking place, and where, could help researchers to identify stressed systems, Kramer says.

The devices use commercially available sensors and are built with the help of 3D printers. Kramer and Kanazawa hope to bring down the cost of the underwater version from its current US\$500 and get it into the hands of as many scientists as possible. "We need an army of people making high-quality measurements," Kramer says.

COMPUTER-ASSISTED VISION

Marine microbiologist Arjun Chennu has developed an underwater imaging system to collect even more detailed data across a greater radiation spectrum. Coral ecologists then annotate the images, and the information is fed into a neural-network algorithm that is based on open-source machine-learning software and is similar to the one developed by Beijbom. The machine's 'hyperspectrum' means that it can capture much more information than can the human eye, says Chennu, who works at the Max Planck Institute for Marine Microbiology in Bremen, Germany. This makes it easier to differentiate between corals that look similar in standard images. "For example, we resolve the often-used 'other coral' categories

into their proper taxonomic types, and also include sponges, macroalgae and seagrass in our predictions," he says.

Others have adapted commercially available software that is already used to map landscapes and analyse landslides by overlapping 2D images into 3D models. PhD student John Burns at the University of Hawaii at Manoa's Institute of Marine Biology uses a program called Agisoft PhotoScan, which costs \$549 for an educational licence for the professional edition. Free software is available, but it is less sophisticated, Burns says.

The models — which can achieve a resolution of just 1 millimetre when used with good cameras — can be analysed by people or computers to identify coral species and quantify reef coverage. But, because they're 3D, they can also be used to track structural changes as reefs bleach and break down owing to high ocean temperatures — a new kind of ecological information.

For Burns, the beauty of the method is its simplicity: data can be collected quickly and with minimal training. "This method just lets you take hundreds of thousands of single-lens images with your camera, and then you are essentially stitching them together," he says.

STANDARDIZED STORE

Sophisticated technologies aren't the only answer, says Emily Darling, a marine ecologist with the Wildlife Conservation Society in New York City. Because separate research efforts are collecting ever-greater quantities of data on coral reefs, it is important that they collect standardized data sets and store them in a repository that can be accessed by the entire community.

In an effort to collect systematic data on the recent global bleaching event, for example, Darling and her colleagues came up with a very simple technology — an Excel spreadsheet that scientists around the world can use to register various data on reef conditions. The value is that when scientists come out of the water, they can immediately import and analyse their data, and Darling now has uniform results from more than 61,000 reef colonies in 13 countries. Roughly 58% showed bleaching.

Ultimately, Darling says, coral ecologists need to converge on some kind of a central repository for the full suite of information that they are collecting around the world. "We need places where data are accessible, where they are telling stories, and where people can go and figure out whether conservation actions are working or not," she says. "We need to be able to answer those questions a lot faster." ■

CORRECTION

The story 'The paper promoters' (*Nature* **536**, 113–114; 2016) should have made clear that Altmetric.com collects data from both mainstream and social media.

CAREERS

BALANCE Five ways to break a work addiction for go.nature.com/2bfstzt

FIX SCIENCE Young researchers draft treatise for radical change go.nature.com/2bhcmzt

NATUREJOBS For the latest career listings and advice www.naturejobs.com

SEMICONDUCTOR



Joe Gerhardt, one-half of the UK artist duo Semiconductor, explores the archives at CERN with archivist Anita Hollier as part of the COLLIDE initiative.

ART-SCIENCE COLLABORATIONS

Change of perspective

Pick up a lump of clay or stare at a Leonardo water drawing — your science, not just your frame of mind, will benefit from it.

BY SHEILA MULROONEY ELDRED

After earning her PhD in Earth and planetary sciences, Johanna Kieniewicz found herself in a coveted tenure-track job. But as she dug more deeply into her work, she felt her field of vision narrowing — and not in a good way. Extreme focus left her worried that she was stifling her creative side.

“With the intensity of those sorts of jobs, it becomes all that you do,” she says. “I was in danger of losing the bigger picture.” To re-engage with her artistic side — she had always had a penchant for drawing and making things with her hands — she took a leave of absence and

went to art school. There, she came to realize how skills taught in the art world could influence science. Asking difficult questions about purpose and ethics, or imagining both fantastic and terrifying futures, helps scientists to put their work in perspective, she says. She used her art experience to nab a dream job as head of outreach and engagement at the Institute of Physics in London, where she coordinates with art museums and theatres to pull the public into conversations about science. “Ultimately, both artists and scientists are asking big questions about the world,” Kieniewicz says. “A lot of rich and exciting stuff is happening between them.”

Although Kieniewicz took her affinity for

art to the far end of the spectrum, attending art school is hardly a prerequisite for those who hope to expand their scientific horizons and frame an experiment differently or get past a sticking point. Even a rudimentary interest in art can help to shift a researcher’s perspective. Routes into the realm include creating your own art, collaborating with artists and viewing art that resonates with you.

Making art can be very helpful for scientists when they are failing to make progress. “Sometimes you have to dive in deeply, but sometimes you’re stuck and have to get unstuck,” says Robbert Dijkgraaf, director of the Institute for Advanced Study in Princeton, New Jersey. ►

► He advises his students to engage in some form of art when they encounter seemingly insurmountable obstacles in their research.

Cancer researcher Silvia Balbo relates to that recommendation. She has access to an art studio for precisely that purpose. It's been her escape ever since she took a sculpting class in high school in Turin, Italy. "Whenever I feel like things are stuck, I go back to it," she says. She made use of the studio many times during a particularly gruelling three-year project on how tobacco smoke and alcohol damage DNA and contribute to cancer.

Balbo would often head, exhausted, directly from her lab at the University of Minnesota in Minneapolis to the studio. "I'd be super tired, but I'd get there and then suddenly feel super energized," she says. "On those days where you feel like you haven't accomplished anything, it's nice to get a feel for making something. I picked clay because it's constructive: all of a sudden, I have a piece. That immediate outcome is very rewarding."

Clay modelling also gives her a chance to turn off the structured, analytical part of her brain, she says, and allow intuition and creativity to take over. Often, she leaves the studio with a fresh outlook on a knotty experiment. "I'll get out of there and realize, 'Oh, I had not thought of it in this way before,'" she says.

Over the course of that project, Balbo sculpted, fired and glazed four pieces: nude women in various languid postures drenched in streaks of blue glaze that she now displays in her home. Ultimately, her team had a breakthrough, and published the findings. In addition to unlocking new ideas in the lab, she credits the sculpting with helping her to stay on track. "It's very energizing to have a peek into the art world and recharge your batteries," she says.

EYE OF THE BEHOLDER

The pay-offs of art involvement need not come just from creation. Simply looking at it can also bring benefits: gazing at other people's creative endeavours can help scientists to find inspiration and come up with new approaches. Chemist Catherine Murphy at the University of Illinois at Urbana-Champaign is drawn to close-ups of natural objects, the bright colours of inorganic compounds and the brilliant hues of gold nanomaterials. She has a copy of artist Georgia O'Keeffe's *Red Poppy No. VI* on her office wall just so that she can stare at the flower's vibrant scarlet petals. She once bought a painting at an art fair that looked to her like proteins seen through an atomic force microscope (not what the artist had in mind, she says). "I thought it was really interesting that the same visual could be perceived in so many different ways," she says. "In science, the more different perspectives you have on the phenomena you're studying, the richer the understanding becomes."

Other ways to stretch scientific thinking are discovered when researchers collaborate with

Scientists who have no experience in art can still find ways to engage with their creative side. "If you have any curiosity for art, I'd encourage you to give it a try," says Silvia Balbo, a cancer researcher at the University of Minnesota in Minneapolis.

- Make a friend in the art department of your institution. Go to a thesis presentation and invite an art student to visit your lab.
- Apply for a residency or offer to participate in one. Here are a few: the Massachusetts Institute of Technology Center for Art, Science & Technology in Cambridge (go.nature.com/2bnppjh), Arts@CERN in Geneva, Switzerland

artists. So effective have these partnerships been for stimulating scientific creativity that some research institutions have established programmes to encourage them (see 'Up your art quotient'). Europe's particle-physics laboratory, CERN, for example, established a programme called COLLIDE to foster ingenuity through the exchange of ideas between scientists and artists. The initiative brings world-class artists to the laboratory and campus in Geneva, Switzerland, for a residency of up to three months.

CERN theoretical physicist Luis Álvarez-Gaumé (who moves to Stony Brook University in New York this month), recently worked with two UK artists as part of the initiative. The artists used scientific data and computer-generated animation to probe how scientific instruments and discoveries in particle physics influence the perception of nature. Explaining his work to them for their upcoming piece helped Álvarez-Gaumé to find holes in his own knowledge. "It allows us to really see, to appreciate and understand what we are talking about," he says. Kieniewicz agrees that working with artists helps scientists to reframe their thinking. "The artist will come in from a bit of a tangent, probing areas where scientists wouldn't think to probe," she says. "They are really good at asking 'what if' questions — 'what if we could hear the Higgs boson?'"

Art-science collaborations can produce other benefits, too. Murphy established a programme at her lab in which university art students come in and ask questions of her chemistry pupils. She quickly realized that her students rapidly improved at communicating their work and ideas. "When you're giving a presentation to a totally non-scientific audience, you have to be able to communicate really well," she says.

RESOURCES

Up your art quotient

(go.nature.com/2b5b9jb), The Guapamacátaro Center for Art and Ecology in Michoacan, Mexico, and the Institute for Advanced Study in Princeton, New Jersey (<http://go.nature.com/2brypgx>).

- Participate in a collaboration such as those organized by the Institute of Physics in London (go.nature.com/2b9ycel).
- Sign up for a drawing or pottery class through community education, audit a class at your university or search for museum-based programmes.
- Search Meetup.com for art-related outings.
- Search Twitter for #sciart. **S.M.E.**

And scientists are often awestruck by seeing artists portray what they've learned in completely new ways. Artists who have worked alongside Murphy's students, for example, have created everything from a dance interpreting the view through an electron microscope to a computer-sized block of canvas with light bulbs shining through at various levels of brightness, inspired by the gold particles that the artist glimpsed through a microscope. Because the results are usually exhibited to both scientists and artists, they provide an ideal opportunity for interdisciplinary conversations.

The collaborations spawn more than impressive art — they are rich for researchers too, says Martin Kemp, an emeritus art historian at Trinity College in Oxford, UK, who specializes in visualization of science and has written a book called *Structural Intuitions: Seeing Shapes in Art and Science* (Univ. Virginia Press, 2016). He says that perception is deeply embedded in the brain by the end of formal schooling, yet researchers must embrace other ways of thinking and visualizing, and can do so through making or viewing art. He thought he was leaving science forever when he went to do graduate studies at an art institute — until he stumbled across Leonardo da Vinci's water drawings. The detailed sketches depicting patterns and shapes of water, wind and air reflect the theory of hydrodynamics, he says — completely applicable to both art and science. "I felt I'd come home," Kemp says.

Although that sort of leap is practical for very few ("It doesn't help to have art school on your CV to get funded," Balbo says), most of the bright scientists Kemp knows engage in the arts in some form. And some even say it is essential to their careers.

"If I had not gone to art school, I don't believe I would be a scientist today," Dijkgraaf says. ■

Sheila Mulrooney Eldred is a freelance writer in Minneapolis, Minnesota.

SKIN HUNGER

A sense of loss.

BY BO BALDER

Kirsten knocked. The carved ebony knocker landed soundlessly on velvet. The door swung open and she entered the dark hallway. The door closed, not with a sound but with an intensification of the air inside, making it dryer, darker and more silent.

She waited until her eyes had got used to the darkness. Sometimes she felt anticipation, sometimes dread. Often she felt nothing much. She appreciated the efforts of her hosts, but gratitude no longer had the power to mitigate the realization that she'd never get out of here. As the only survivor of the crash, what she missed most was other humans.

A light glowed up at the end of the hallway. The layout of the facility changed with every visit. Surprise and curiosity were emotions as well, and again, they did their best.

Kirsten started walking, reluctant to let go of her anticipatory state. Her feet touched warm, non-resonant flooring that came close to being wood. Close enough for her soles, if not for her ears and sense of smell. Even though her mind was sinking into despair, her body could still react to small pleasures. She rolled her soles over the floor, savouring the moment, wishing she could store the sensation.

A seven-foot dragon burst from the ceiling and shrieked in her face.

Kirsten's heartbeat accelerated slightly. If only she could still experience the full range of emotions. Her mind knew this was a side effect of depression, but that didn't seem to change anything. The dragon flapped its kimono wings and stomped its booted feet. The bright reds and yellows of the dress, the gyrations and dips of the dance were gorgeous. As were the tinkling, winking head-dress decorations, like seahorse antennae. Kirsten clapped politely. The unmoving mask turned in her direction, the figure took a bow and vanished.

Beauty and grace moved her only a little. Her hosts had tried ugliness and stench

and danger to stir her, and that had worked in the beginning. But everything palled eventually, so they had reverted to recreating familiarity. Although Noh hadn't been in her cultural vocabulary back on Earth.

Fuzzy blackness descended over her head. Kirsten fought, from real surprise, but her captors jostled her, rolled her over, carried

The aliens lowered Kirsten into a warm mud bath, which was a regular occurrence during the ritual. It was the one thing they got right, simply because she'd never had one back home. Submerged younglings massaged her sore body with their tentacles.

Something warm touched her shoulders.

She leaned forward so they could massage her. Warm, slightly rough fingers dug into her tight shoulders in just the right places. How had they learned this? She imagined tentacled beings practising on each other's non-existent shoulder muscles, or on Kirsten dolls. She smiled. That was a good feeling, even if her smile muscles tired quickly.

Then a sensation ghosted over her back, as if the masseur's hairy underarm accidentally touched her back. Her body flooded with adrenaline, the hair rose on her neck, her breath caught in her throat. The touch of another human being's skin. The thing she'd

missed more than words.

"Who are you?" the real Kirsten said, sharply.

She must have startled the masseur, because she glimpsed a flash of an old cashmere sweater of hers it held in two tentacles. The illusion shattered.

She cried so hard it hurt. For that one moment she'd been back to her old self again, but it was worse than feeling everything through a veil. She was going to die alone. Without ever having touched another human being again.

She rose from the mud-bath, drying her tears. No more rituals. She never wanted to experience another moment of hope like that. It hurt too much. ■

Bo Balder is the first Dutch author to be published in Fantasy & Science Fiction. Her short fiction has appeared in *Crossed Genres*, *Futuristica* Volume 1 and other venues. Her SF novel *The Wan* is published by *Pink Narcissus Press*.



ILLUSTRATION BY JACEY

her somewhere while she bounced on their bodies, was banged against obstacles and felt sick to her stomach. Surprise? Check. Discomfort? Check. Nausea? Bruising? Even unpleasant sensations were welcome.

She was squeezed out of her sack, the clothes ripped from her body, rolled around in scalding water and beaten with something like rubber hose. Only it wouldn't be rubber, or hose. Probably her hosts' tentacles. Just being aware of that possibility made it feel less real.

Gentle touches soothed her skin with water almost smelling like jasmine. So close. How did her nose even know it wasn't the real thing, after so long without it? Maybe she was hallucinating and that was why nothing felt quite real. It was as if the real Kirsten, the one with depth and warmth, was just behind a curtain, tantalizingly within reach. She kept thinking that she'd find that person again, soon. Tomorrow.

ON NATURE.COM

Follow Futures:

@NatureFutures

go.nature.com/mtoodm